Introdução às Análises Metagenômicas

Prof. Dr. Daniel Guariz Pinheiro

Departamento de Tecnologia

Faculdade de Ciências Agrárias e Veterinárias (FCAV)

Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP)

Cultivo de microrganismos

- Atualmente menos de 1% de todos os microrganismos existentes no mundo podem ser cultiváveis no laboratório.

- Fenômeno limitante para a compreensão da fisiologia microbiana, genética e a ecologia das comunidades

Caracterização morfológica, fisiológica/bioquímica.

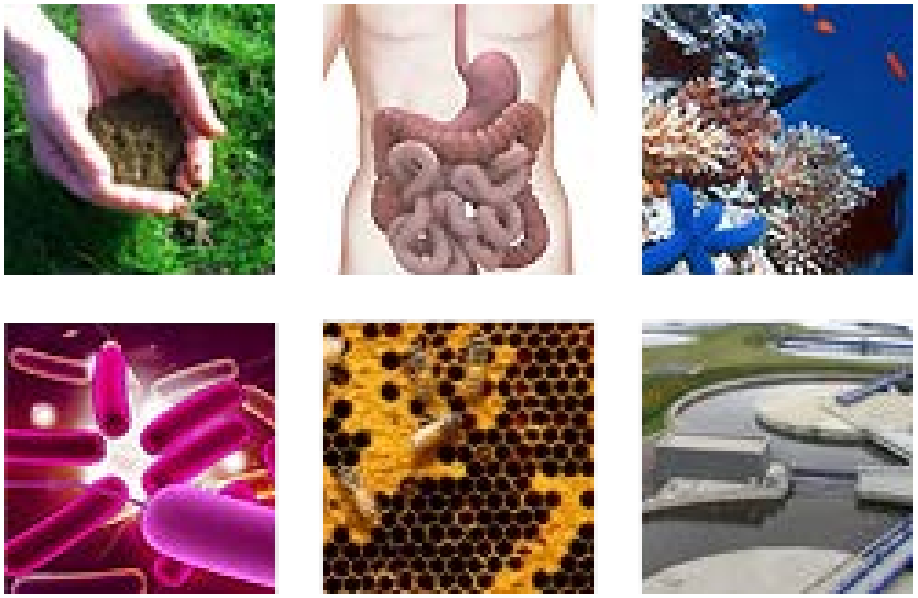
Agar Tríptico de Soja (TSA)

Morphological diversity typical of microorganisms cultured from soil on a broad spectrum medium, tryptic soy agar.

Many lines of evidence show that fewer than 0.1% of the microorganisms in soil are readily cultured using current techniques *(Handelsman et al., 1998)*

Metagenômica

- Estudo de todos os genomas presentes em uma amostra ambiental. Sem a necessidade de isolamento, cultivo ou identificação.



Definição: Metagenômica

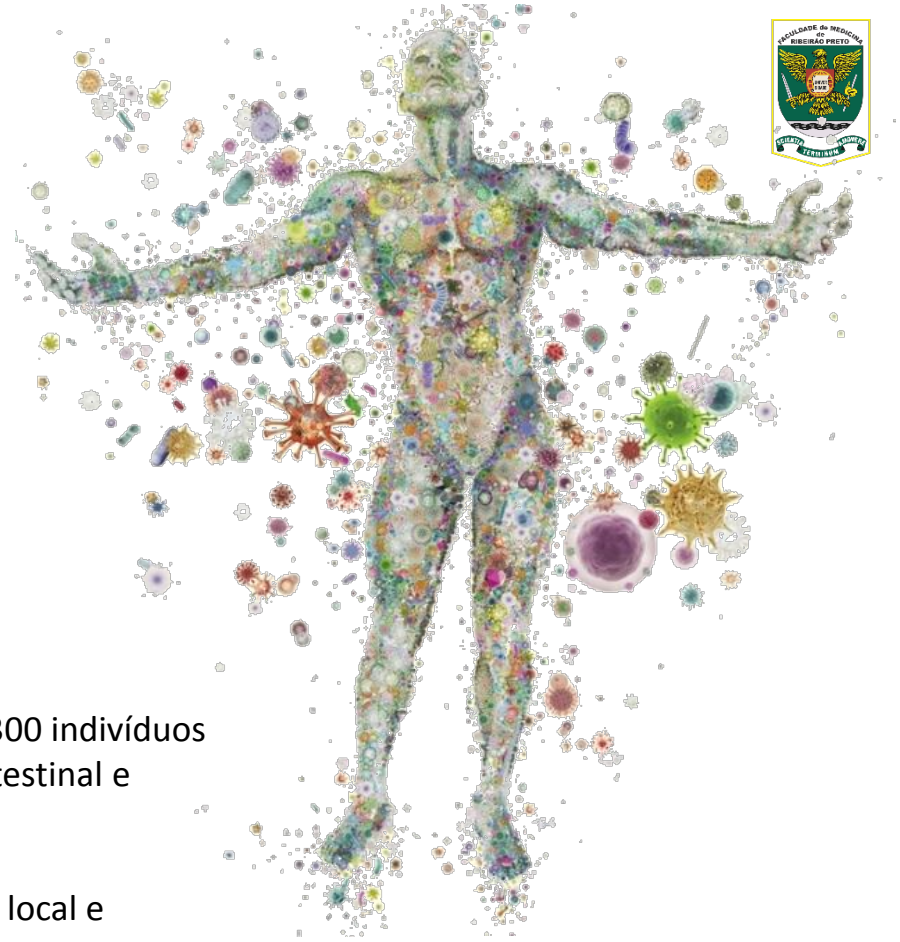
A new frontier of science is emerging that unites biology and chemistry — the exploration of natural products from previously uncultured soil microorganisms. The approach involves directly accessing the genomes of soil organisms that cannot be, or have not been, cultured by isolating their DNA, cloning it into culturable organisms and screening the resultant clones for the production of new chemicals.

(Handelsman et al., 1998)

- *(also referred to as environmental and community genomics)*
- *… is the genomic analysis of microorganisms by direct extraction and cloning of DNA from an assemblage of microorganisms.*
  (Handelsman, 2004)

Microbioma Humano

- Nosso outro genoma



HMP1

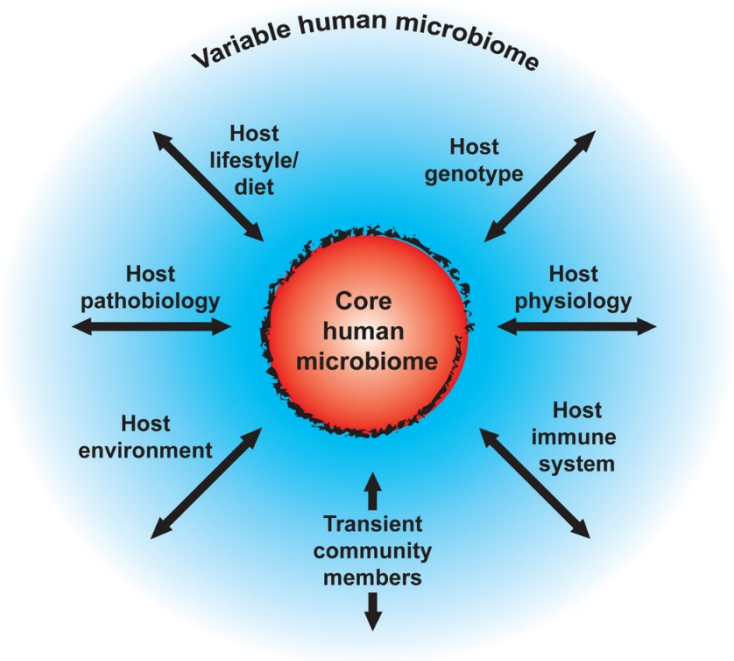
- 2008 a 2013
- Caracterização das comunidades microbianas a partir de 300 indivíduos saudáveis em diferentes locais do corpo humano: trato intestinal e urogenital, cavidade oral, pele, etc.
- Sequenciamento de rRNA 16S
  - Caracterização da comunidade microbiana em cada local e identificação do “core microbiome”;
- Sequenciamento de DNA Total (WGS – Whole Genome Shotgun)Metagenomic whole genome shotgun (wgs)
  - Genes e vias biológicas
- 14.23 terabytes de dados
- Broad Institute, the Baylor College of Medicine, Washington University School of Medicine, and the J. Craig Venter Institute, the Data Analysis and Coordination Center (DACC), e muitos investigadores

NIH HUMAN MICROBIOME PROJECT

'human supraorganism'

The microbes that live inside and on us (the microbiota) outnumber our somatic and germ cells by an estimated 10-fold.

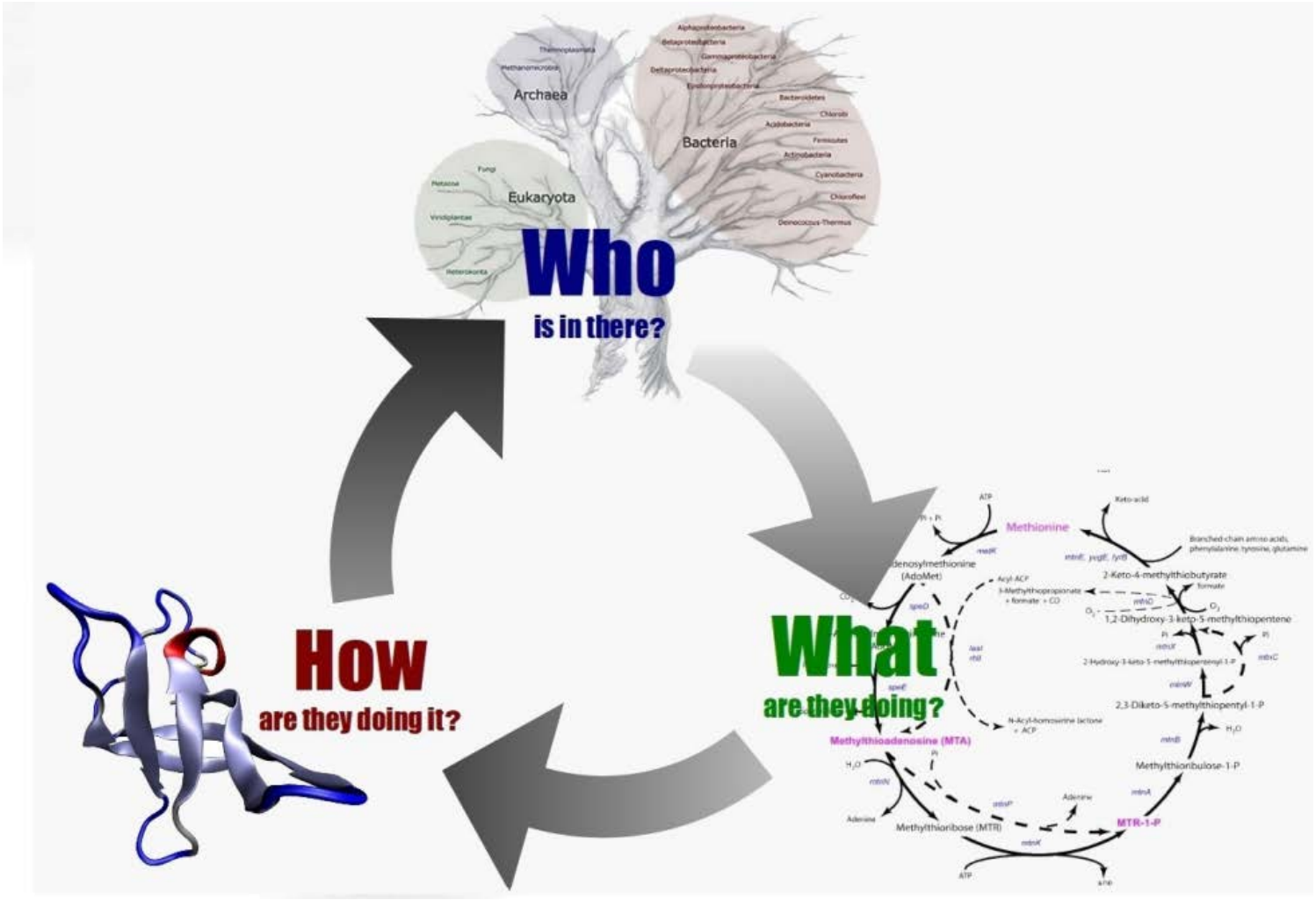
… a composite of microbial and human species



iHMP

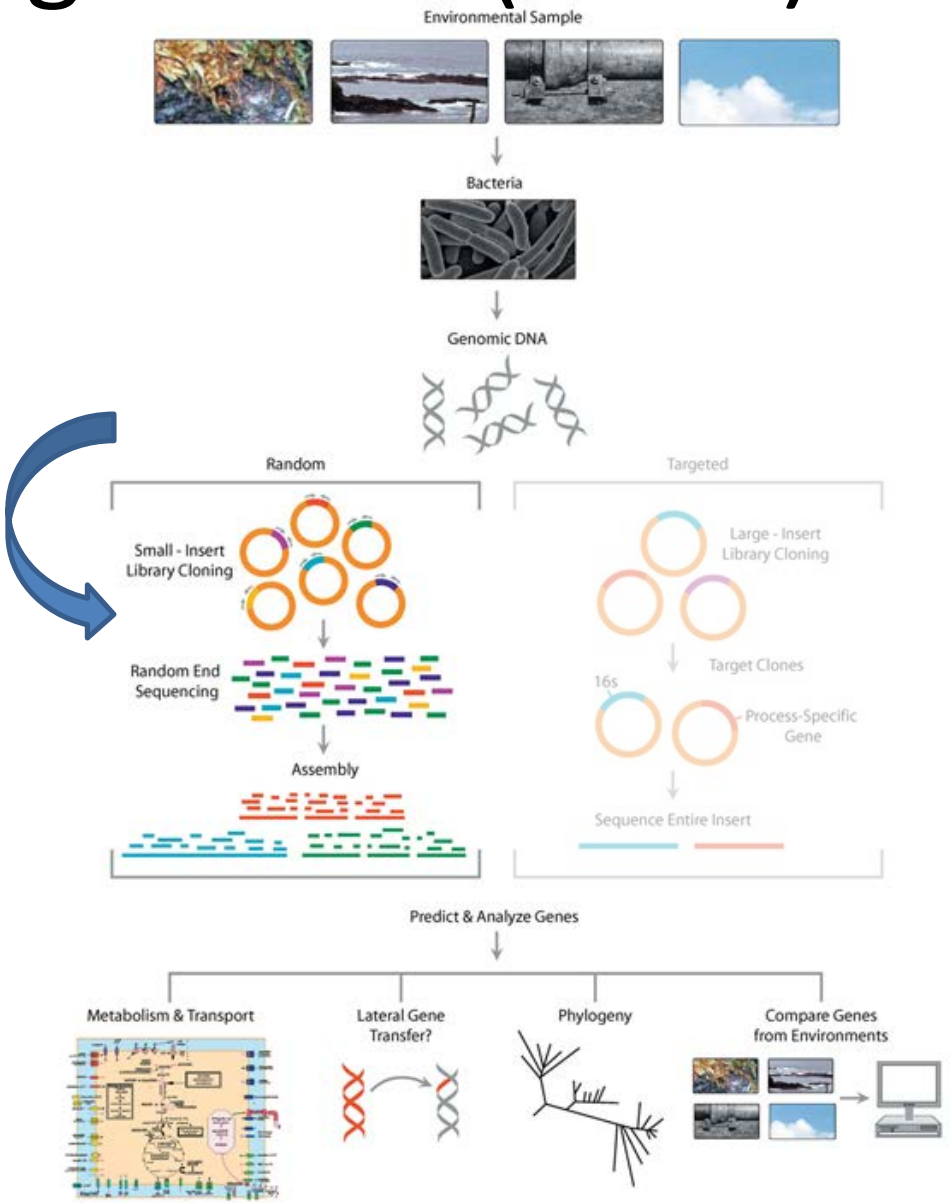
- NIH Integrative Human Microbiome Project
- Estudos de coorte
- O objetivo desta segunda fase é gerar recursos que possam contribuir para a caracterização da microbiota humana para posterior compreensão de qual é o impacto do microbioma na saúde humana e nas doenças.

Principais questões



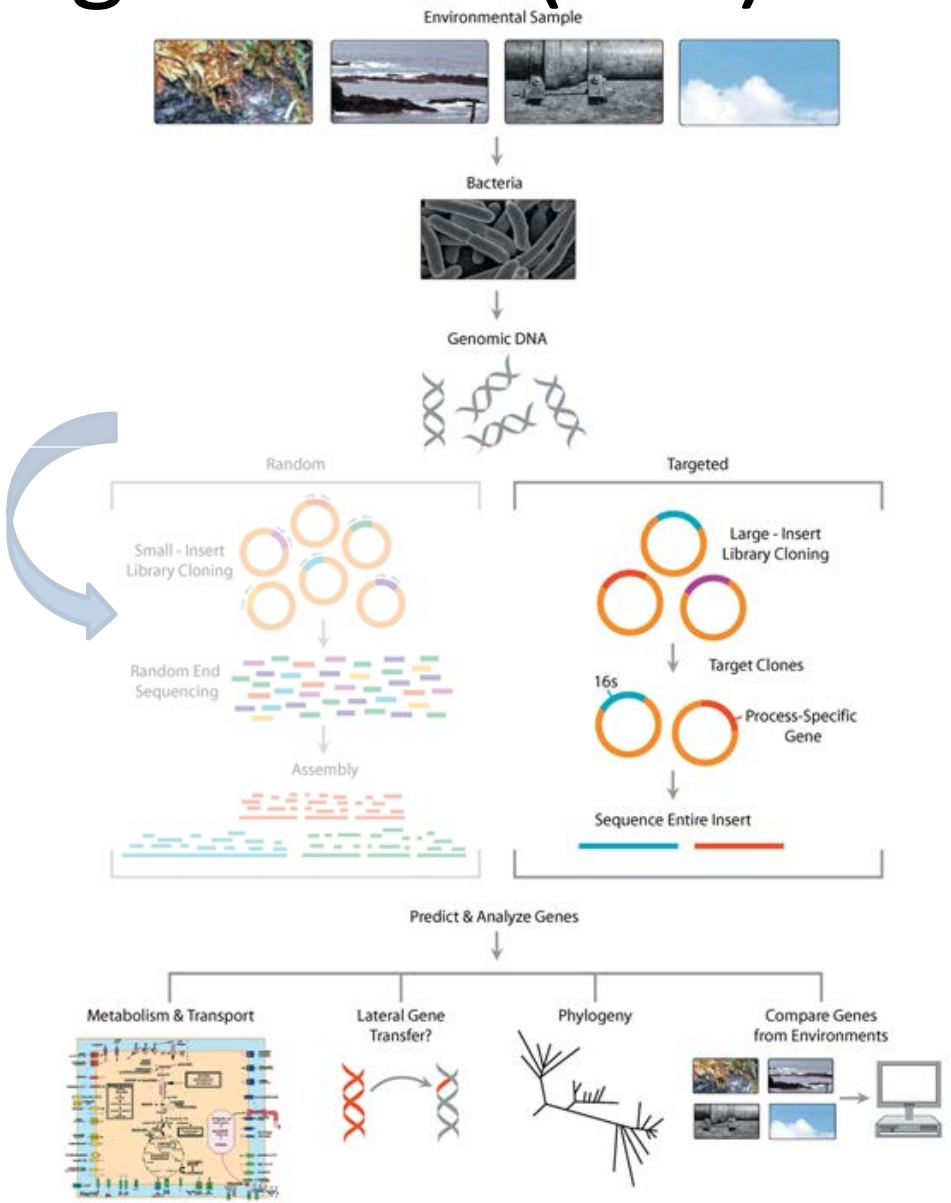
Dados de metagenomas (WMS)

- Estratégia *shotgun* de sequenciamento de **DNA total**
  - *Whole Metagenome Sequencing*
  - Alternativa para estudo da microbiota não cultivável
    - Permitindo investigar os seguintes aspectos
      - **Quem** está lá?
      - **Quantos** estão lá?
      - **O que** são capazes de fazer?



Dados de metagenomas (TAS)

- Estratégia de sequenciamento de ***Amplicons* (Alvos)**
  - ***Targeted Amplicon Sequencing***
  - Alternativa para estudo da microbiota não cultivável
    - Permitindo investigar os seguintes aspectos
      - **Quem** está lá? (principalmente)
      - **Quantos** estão lá?
      - **O que** são capazes de fazer?



Single-cell sequencing

nature REVIEWS GENETICS

nature.com ▸ journal home ▸ archive ▸ issue ▸ progress ▸ abstract

ARTICLE PREVIEW

view full access options ▸

NATURE REVIEWS GENETICS | PROGRESS

ARTICLE SERIES: Applications of next-generation sequencing

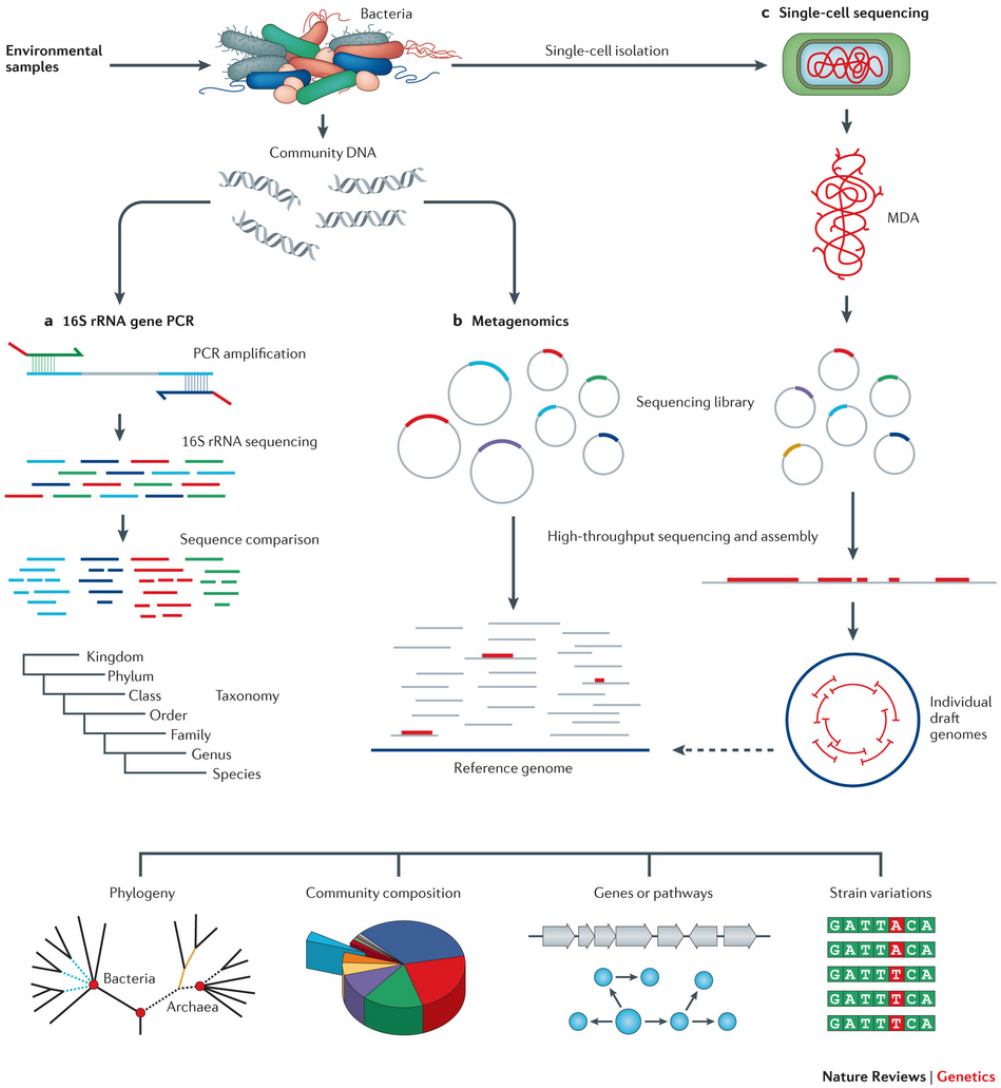
Recent advances in genomic DNA sequencing of microbial species from single cells

Roger S. Lasken & Jeffrey S. McLean

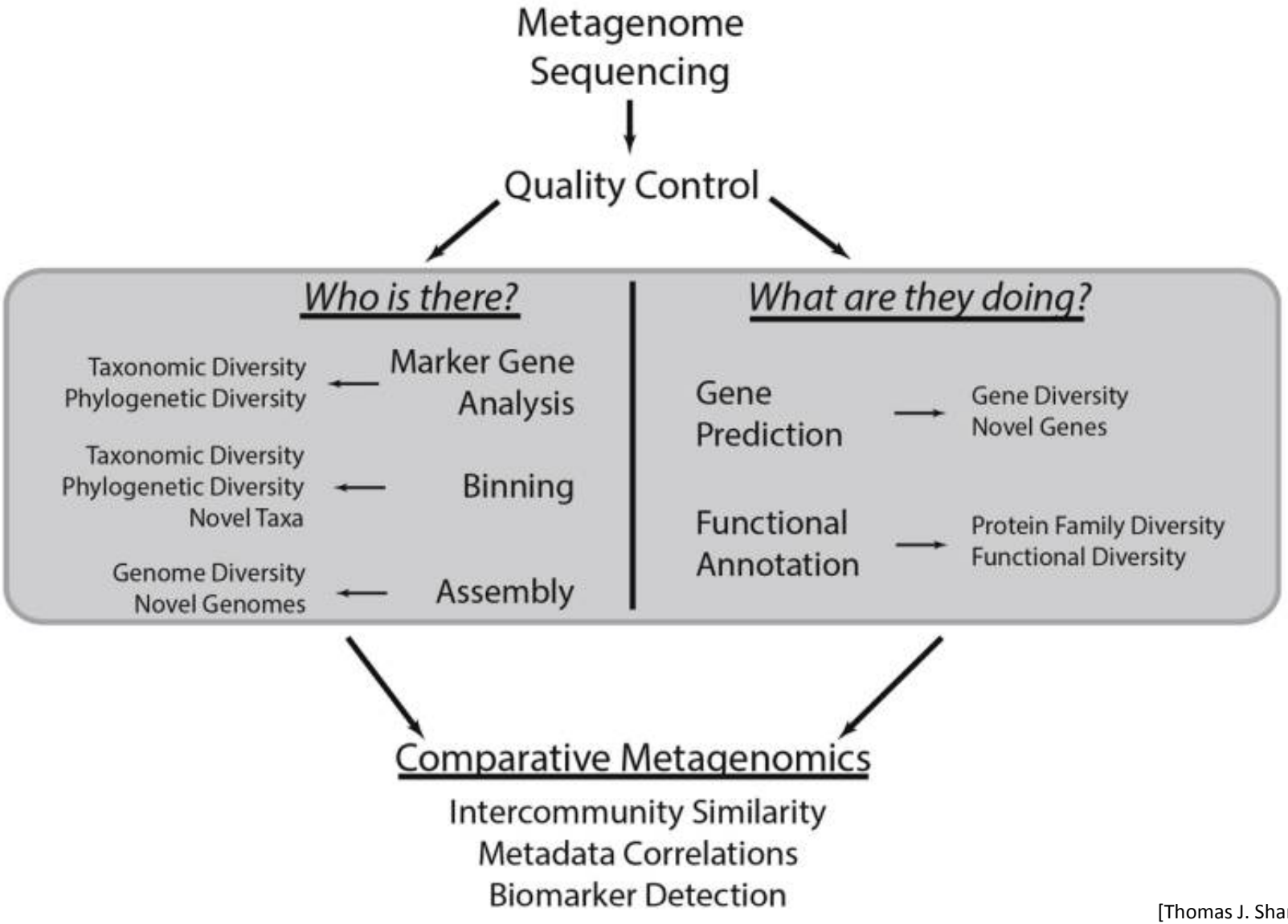
Affiliations | Corresponding author

Nature Reviews Genetics **15**, 577–584 (2014) | doi:10.1038/nrg3785

Published online 05 August 2014



Análises metagenômicas

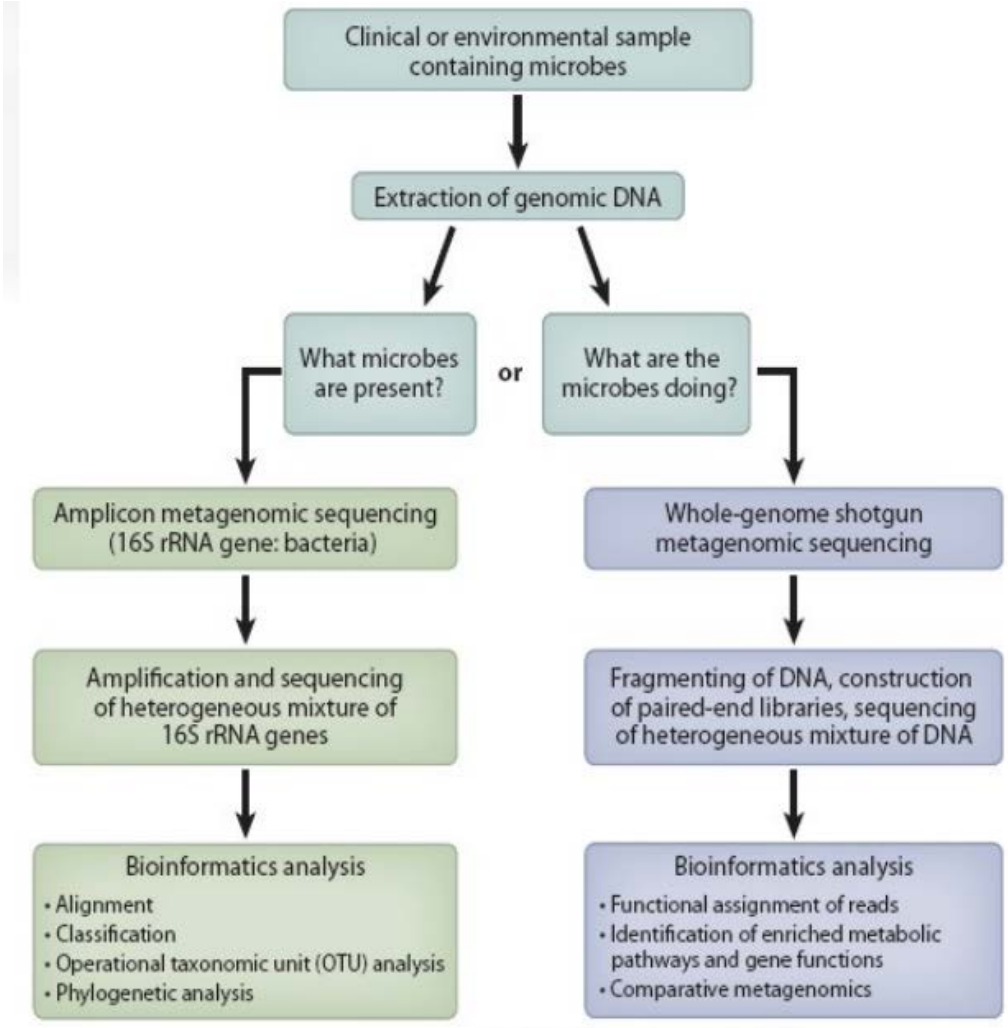


[Thomas J. Sharpton, 2014]

Desafios

- Dados são complexos e enormes;
  - Difícil determinar de que genoma determinada leitura teve origem;
  - Muitas comunidades de microrganismos são muito diversas e a maior parte dos genomas não é completamente representada por uma leitura;
    - Até mesmo **um único gene pode não ser completamente amostrado**, pois as leituras normalmente são mais curtas, e sendo assim não há sobreposição para sua reconstrução completa;
    - Quando há sobreposição entre leituras, ainda há a possibilidade disso conduzir a **erros no alinhamento ou na montagem** de uma sequência consenso para um único genoma de forma acurada;
  - Em busca de amostragem para representação dos genomas há um **aumento** da **quantidade de dados**;
  - Em especial no caso de microbiotas há a presença de **material genético não desejado do hospedeiro**, o qual pode se sobrepor ao do DNA microbiano (Há métodos de Biologia Molecular para o enriquecimento de DNA microbiano – ex. baseados na diferença de densidade de metilação de ilhas CpG);
    - Plantas com seus genomas enormes tornam esse desafio ainda maior (há estudos que obtiveram DNA metagenômico de filosfera utilizando separação após centrifugação na presença de Percoll - Delmotte et al., 2009);
  - Amostras ambientais estão sujeitas a **contaminações** diversas, uma vez retiradas do ambiente de origem;

Etapas



Grice E.A. & Segre J.A. (2012) The Human Microbiome: Our Second Genome, Annu. Rev. Genomics Human Genet. 13, 151-170

Desenho experimental

Coleta

Extração de DNA genômico

Controle de qualidade das amostras

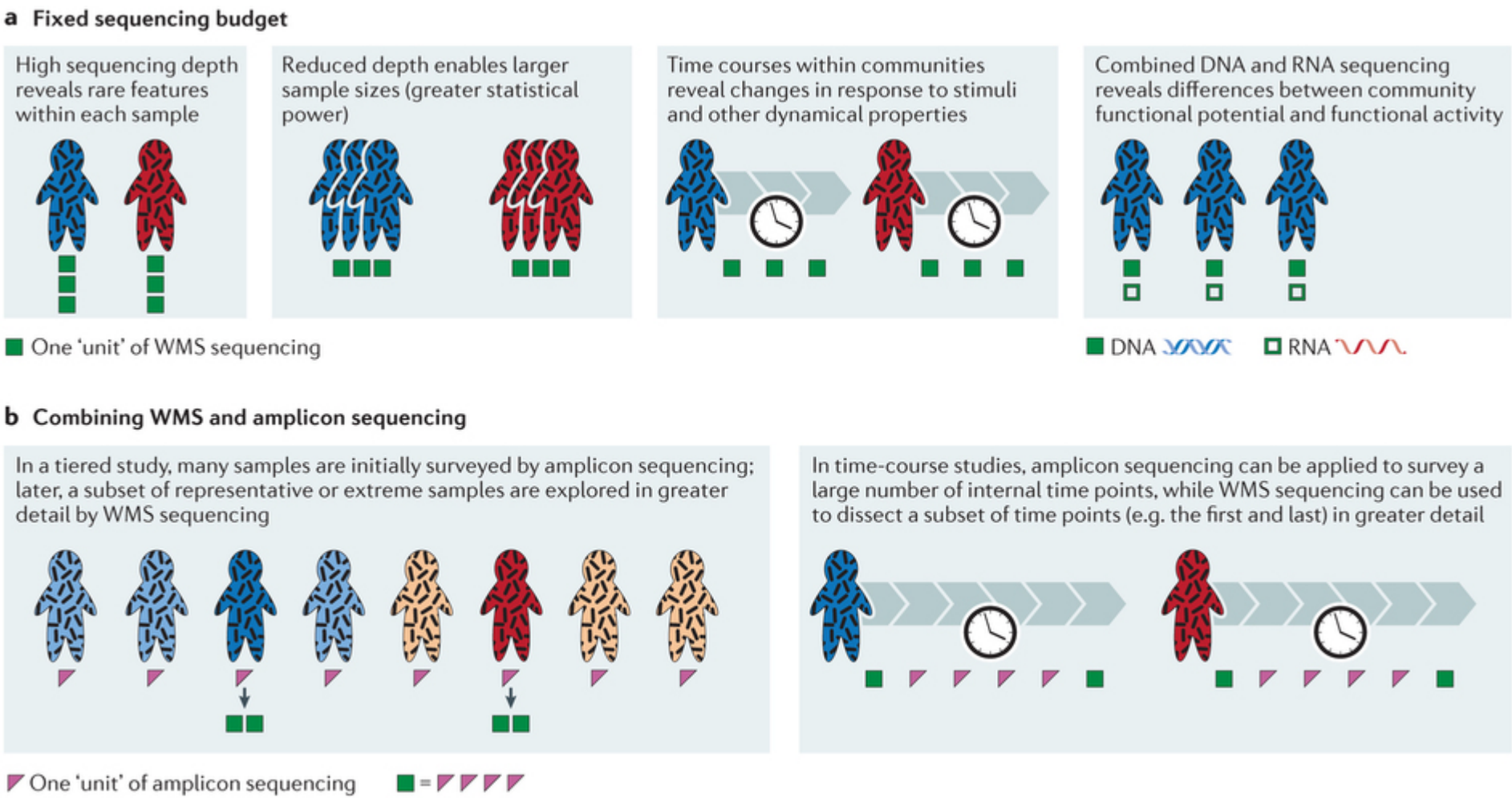
Construção da biblioteca e Sequenciamento

Controle de qualidade dos resultados de sequencimento

Análise

Interpretação biológica

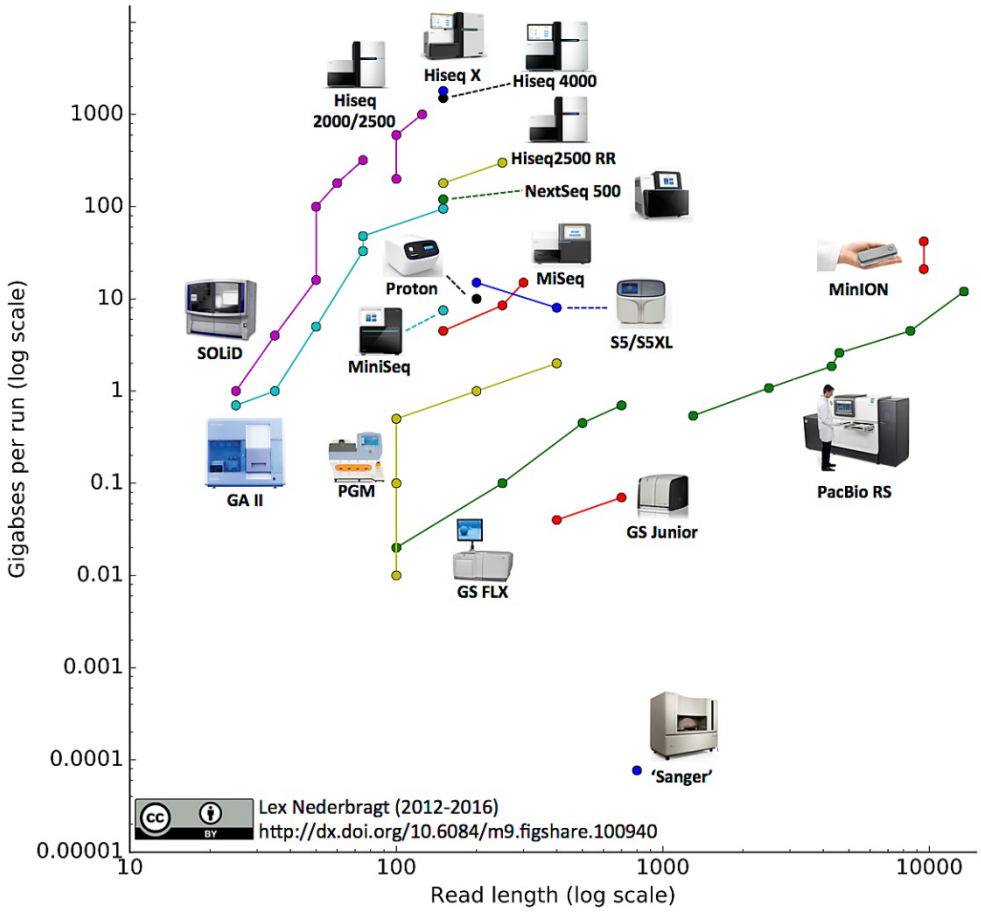
Desenho experimental



http://www.nature.com/nrmicro/journal/v13/n6/full/nrmicro3451.html

Escolha da estratégia de sequenciamento

- Depende da realidade de cada projeto com respeito a:
  - Custo
  - Rendimento
  - Tamanho das leituras (*reads*)
  - Qualidade das leituras



Taxa de erros

Platform-Dependent Error Rates

Table 1 Insertion/deletion and substitution errors on read level for benchtop NGS platforms

Platform	Sequencing kit	Library	Strain	Date of sequencing	Indels per 100 bp	Indels per read	Substitutions per 100 bp	Substitutions per read
GSJ	GSJ Titanium	Nebulization / AMPure XP	Sakai	June 2012	0.4011	1.8351	0.0543	0.2484
MiSeq	2 × 150-bp PE	Nextera	Sakai	June 2012	0.0009	0.0013	0.0921	0.1318
MiSeq	2 × 250-bp PE	Nextera	Sakai	September 2012	0.0009	0.0018	0.0940	0.2033
PGM	100 bp	Bioruptor / Ion Fragment Library	Sakai	July 2011	0.3520	0.3878	0.0929	0.1024
PGM	200 bp	Ion Xpress Plus Fragment	Sakai	July 2012	0.3955	0.6811	0.0303	0.0521
PGM	300 bp	Ion Xpress Plus Fragment	Sakai	August 2012	0.7054	1.4457	0.0861	0.1765
PGM	400 bp[a]	Ion Xpress Plus Fragment	Sakai	November 2012	0.6722	1.8726	0.0790	0.2202

Error rates were calculated by counting indels and substitutions in the mapping against the EHEC Sakai reference sequence for each uniquely mapped read.
[a]Kit was not officially available during time of study.

- ➞ 4-dye cyclic reverse termination approaches (Illumina) are prone to substitution errors
- ➞ 1-dye cyclic no termination approaches (454, Ion Torrent) are prone to indel errors

martin.hartmann@wsl.ch
www.microbiome.ch

Juenemann et al. 2013 (Nat Biotechnol), Quail et al. 2012 (BMC Genomics), Loman et al. (Nat Biotechnol)

Contaminantes...

BioMed Central

BMC Biology

HOME ABOUT ARTICLES SUBMISSION GUIDELINES

RESEARCH ARTICLE | OPEN ACCESS

Reagent and laboratory contamination can critically impact sequence-based microbiome analyses

Susannah J Salter ✉, Michael J Cox, Elena M Turek, Szymon T Calus, William O Cookson, Miriam F Moffatt, Paul Turner, Julian Parkhill, Nicholas J Loman and Alan W Walker ✉

BMC Biology 2014 **12**:87 | **DOI:** 10.1186/s12915-014-0087-z

Received: 15 July 2014 | **Accepted:** 13 October 2014 | **Published:** 12 November 2014

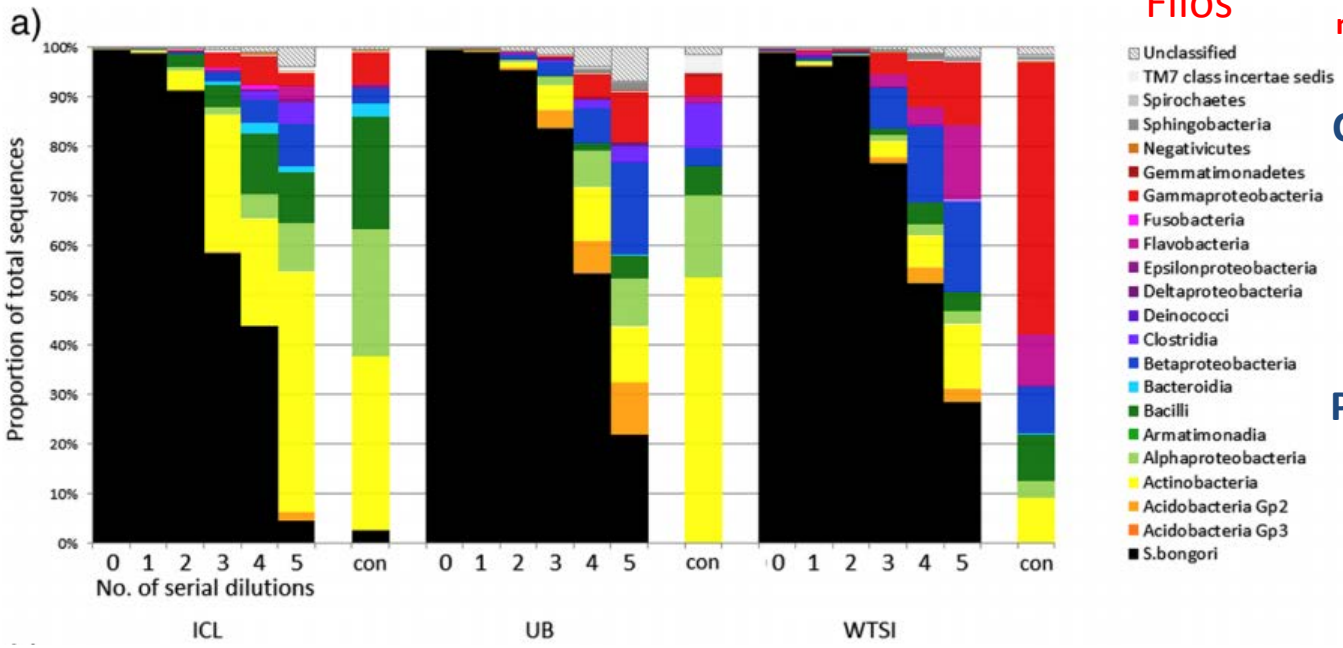
Contaminantes em água e kits

Table 1
List of contaminant genera detected in sequenced negative `blank' controls

Phylum	List of constituent contaminant genera
Proteobacteria	Alpha-proteobacteria:
	Afipia, Aquabacterium[e], *Asticcacaulis, Aurantimonas, Beijerinckia, Bosea, Bradyrhizobium*[d], *Brevundimonas*[c], *Caulobacter, Craurococcus, Devosia, Hoeflea*[e], *Mesorhizobium, Methylobacterium*[c], *Novosphingobium, Ochrobactrum, Paracoccus, Pedomicrobium, Phyllobacterium*[e], *Rhizobium*[c,d], *Roseomonas, Sphingobium, Sphingomonas*[c,d,e], *Sphingopyxis*
	Beta-proteobacteria:
	Acidovorax[c,e], *Azoarcus*[e], *Azospira, Burkholderia*[d], *Comamonas*[c], *Cupriavidus*[c], *Curvibacter, Delftia*[e], *Duganella*[a], *Herbaspirillum*[a,c], *Janthinobacterium*[e], *Kingella, Leptothrix*[a], *Limnobacter*[e], *Massilia*[c], *Methylophilus, Methyloversatilis*[e], *Oxalobacter, Pelomonas, Polaromonas*[e], *Ralstonia*[b,c,d,e], *Schlegelella, Sulfuritalea, Undibacterium*[e], *Variovorax*
	Gamma-proteobacteria:
	Acinetobacter[a,d,c], *Enhydrobacter, Enterobacter, Escherichia*[a,c,d,e], *Nevskia*[e], *Pseudomonas*[b,d,e], *Pseudoxanthomonas, Psychrobacter, Stenotrophomonas*[a,b,c,d,e], *Xanthomonas*[b]
Actinobacteria	*Aeromicrobium, Arthrobacter, Beutenbergia, Brevibacterium, Corynebacterium, Curtobacterium, Dietzia, Geodermatophilus, Janibacter, Kocuria, Microbacterium, Micrococcus, Microlunatus, Patulibacter, Propionibacterium*[e], *Rhodococcus, Tsukamurella*
Firmicutes	*Abiotrophia, Bacillus*[b], *Brevibacillus, Brochothrix, Facklamia, Paenibacillus, Streptococcus*
Bacteroidetes	*Chryseobacterium, Dyadobacter, Flavobacterium*[d], *Hydrotalea, Niastella, Olivibacter, Pedobacter, Wautersiella*
Deinococcus-Thermus	*Deinococcus*
Acidobacteria	Predominantly unclassified Acidobacteria Gp2 organisms

(SALTER et al., 2014). DOI: 10.1186/s12915-14-0087-z.

Experimento para detecção de contaminantes



nitrificantes provenientes dos tanques de armazenamento de água ultra pura (nitrogênio no ar dos tanques)

Com *Salmonella bongori* (preto)

Perfil observado em 20 e 40 ciclos da PCR...

- Amostras referentes aos laboratórios ICL, UB e WTSI.
- DNA não diluído possuí **10^8 células** e a 5ª diluição **10^3 células** (***S. bongori)***.

(SALTER et al., 2014). DOI: 10.1186/s12915-14-0087-z.

Impacto menor: amostras de fezes (> biomassa)

Impacto maior: amostras de sangue e pulmão (< biomassa)

Soluções

- Remoção de sequências indesejadas: Archaea, Chloroplasto de plantas, Chloroplasto de Cianobactérias, outros ...

- Padronização de pessoal responsável pelo uso de kits?

- Controle de sequenciamento (perdemos espaço para adicionar amostras ...);

- Usar DeconSeq v.0.4.3 para remoção de contaminantes (**Boa prática em Bioinformática!!!**).

https://sourceforge.net/projects/deconseq/files/

DeconSeq

PLoS One. 2011 Mar 9;6(3):e17288. doi: 10.1371/journal.pone.0017288.

Fast identification and removal of sequence contamination from genomic and metagenomic datasets.

Schmieder R[1], Edwards R.

Author information

Abstract

High-throughput sequencing technologies have strongly impacted microbiology, providing a rapid and cost-effective way of generating draft genomes and exploring microbial diversity. However, sequences obtained from impure nucleic acid preparations may contain DNA from sources other than the sample. Those sequence contaminations are a serious concern to the quality of the data used for downstream analysis, causing misassembly of sequence contigs and erroneous conclusions. Therefore, the removal of sequence contaminants is a necessary and required step for all sequencing projects. We developed DeconSeq, a robust framework for the rapid, automated identification and removal of sequence contamination in longer-read datasets (150 bp mean read length). DeconSeq is publicly available as standalone and web-based versions. The results can be exported for subsequent analysis, and the databases used for the web-based version are automatically updated on a regular basis. DeconSeq categorizes possible contamination sequences, eliminates redundant hits with higher similarity to non-contaminant genomes, and provides graphical visualizations of the alignment results and classifications. Using DeconSeq, we conducted an analysis of possible human DNA contamination in 202 previously published microbial and viral metagenomes and found possible contamination in 145 (72%) metagenomes with as high as 64% contaminating sequences. This new framework allows scientists to automatically detect and efficiently remove unwanted sequence contamination from their datasets while eliminating critical limitations of current methods. DeconSeq's web interface is simple and user-friendly. The standalone version allows offline analysis and integration into existing data processing pipelines. DeconSeq's results reveal whether the sequencing experiment has succeeded, whether the correct sample was sequenced, and whether the sample contains any sequence contamination from DNA preparation or host. In addition, the analysis of 202 metagenomes demonstrated significant contamination of the non-human associated metagenomes, suggesting that this method is appropriate for screening all metagenomes. DeconSeq is available at http://deconseq.sourceforge.net/.

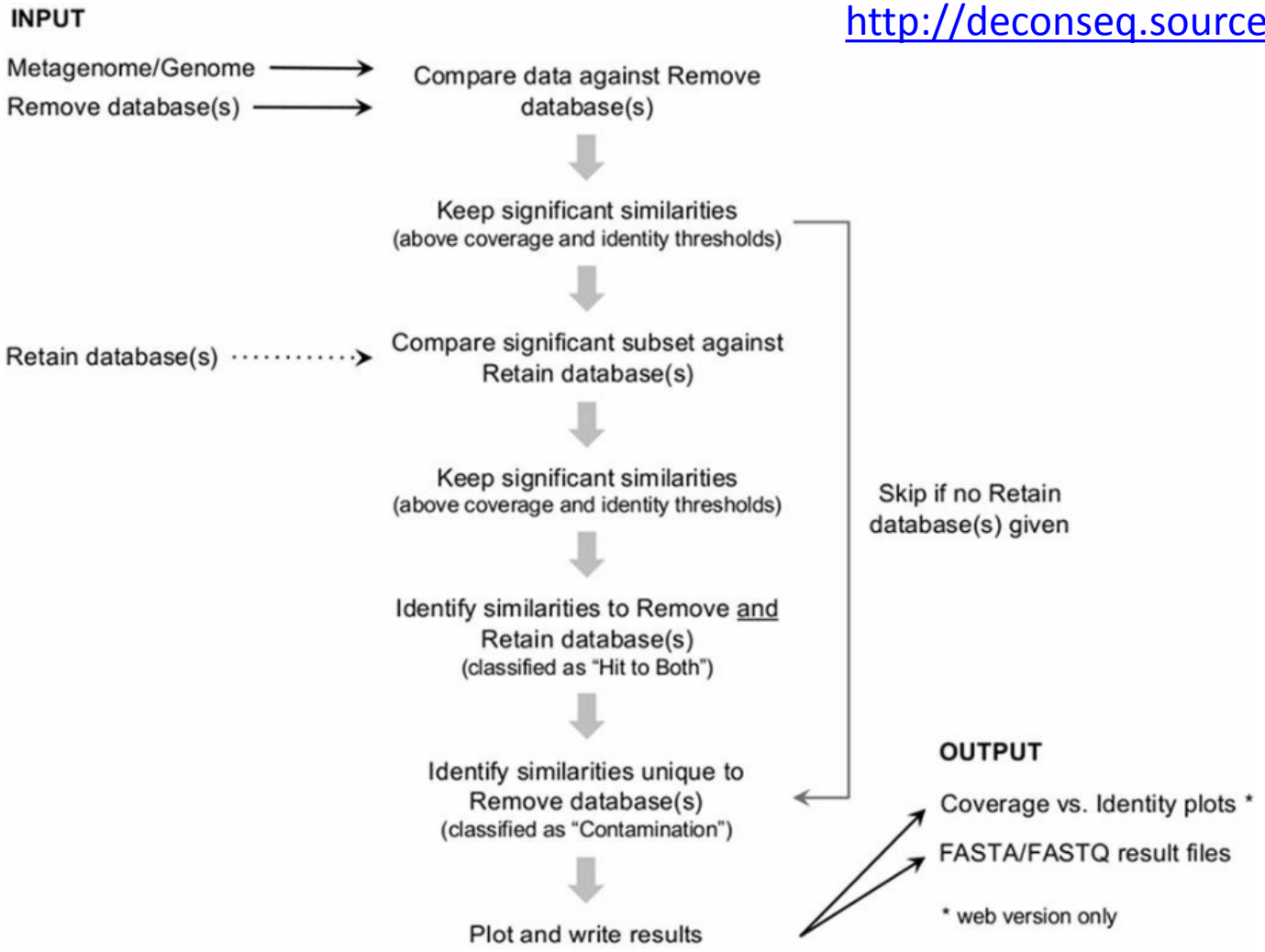
PMID: 21408061 PMCID: PMC3052304 DOI: 10.1371/journal.pone.0017288

[PubMed - indexed for MEDLINE] Free PMC Article

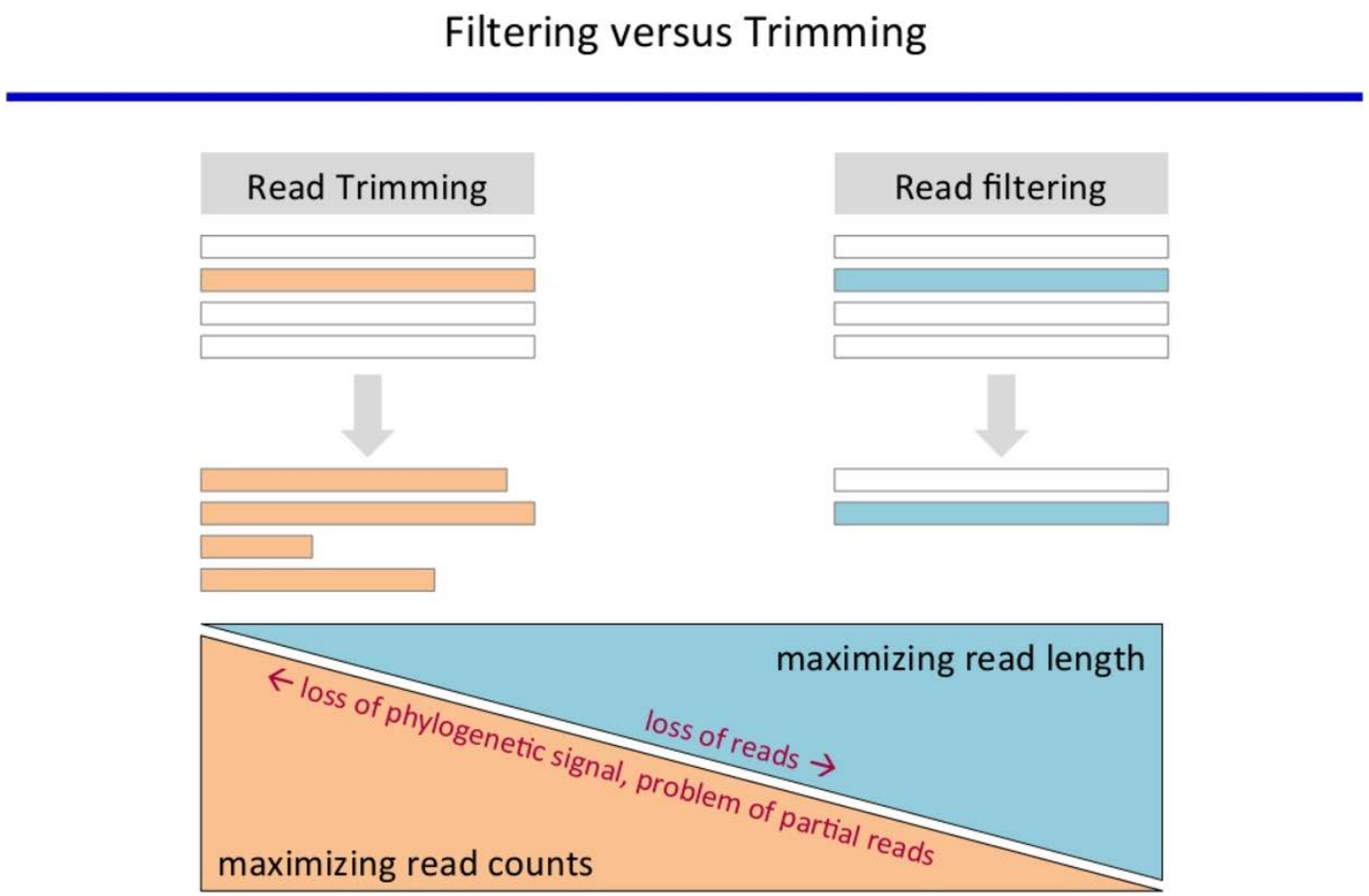
DeconSeq

DeconSeq

http://deconseq.sourceforge.net



Controle de qualidade dos dados de sequenciamento



martin.hartmann@wsl.ch
www.microbiome.ch

PRINSEQ
(Poda e Filtragem - Qualidade)

Bioinformatics. 2011 Mar 15;27(6):863-4. doi: 10.1093/bioinformatics/btr026. Epub 2011 Jan 28.

Quality control and preprocessing of metagenomic datasets.

Schmieder R[1], Edwards R.

⊕ Author information

Abstract

SUMMARY: Here, we present PRINSEQ for easy and rapid quality control and data preprocessing of genomic and metagenomic datasets. Summary statistics of FASTA (and QUAL) or FASTQ files are generated in tabular and graphical form and sequences can be filtered, reformatted and trimmed by a variety of options to improve downstream analysis.

AVAILABILITY AND IMPLEMENTATION: This open-source application was implemented in Perl and can be used as a stand alone version or accessed online through a user-friendly web interface. The source code, user help and additional information are available at http://prinseq.sourceforge.net/.

PMID: 21278185 PMCID: PMC3051327 DOI: 10.1093/bioinformatics/btr026

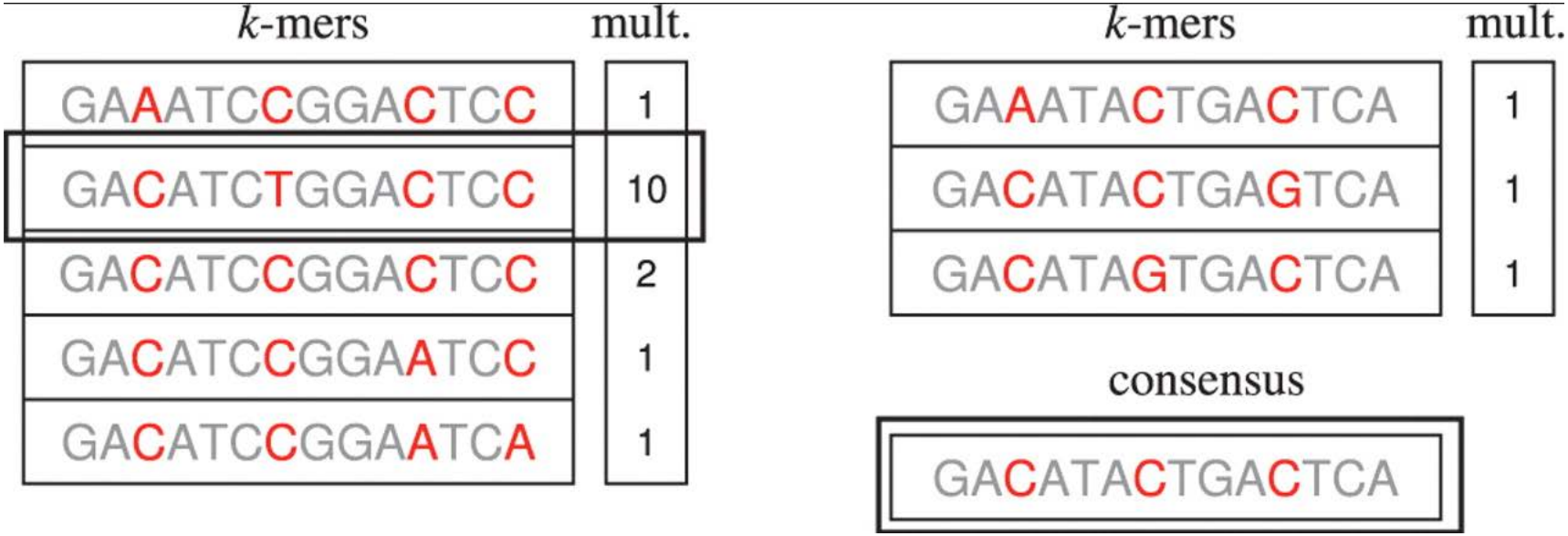
[Indexed for MEDLINE] Free PMC Article

Correção de erros

Error correction of high-throughput sequencing datasets with non-uniform coverage

Paul Medvedev ✉, Eric Scott, Boyko Kakaradov, Pavel Pevzner

Bioinformatics (2011) 27 (13): i137-i141.
DOI: https://doi.org/10.1093/bioinformatics/btr208
Published: 14 June 2011

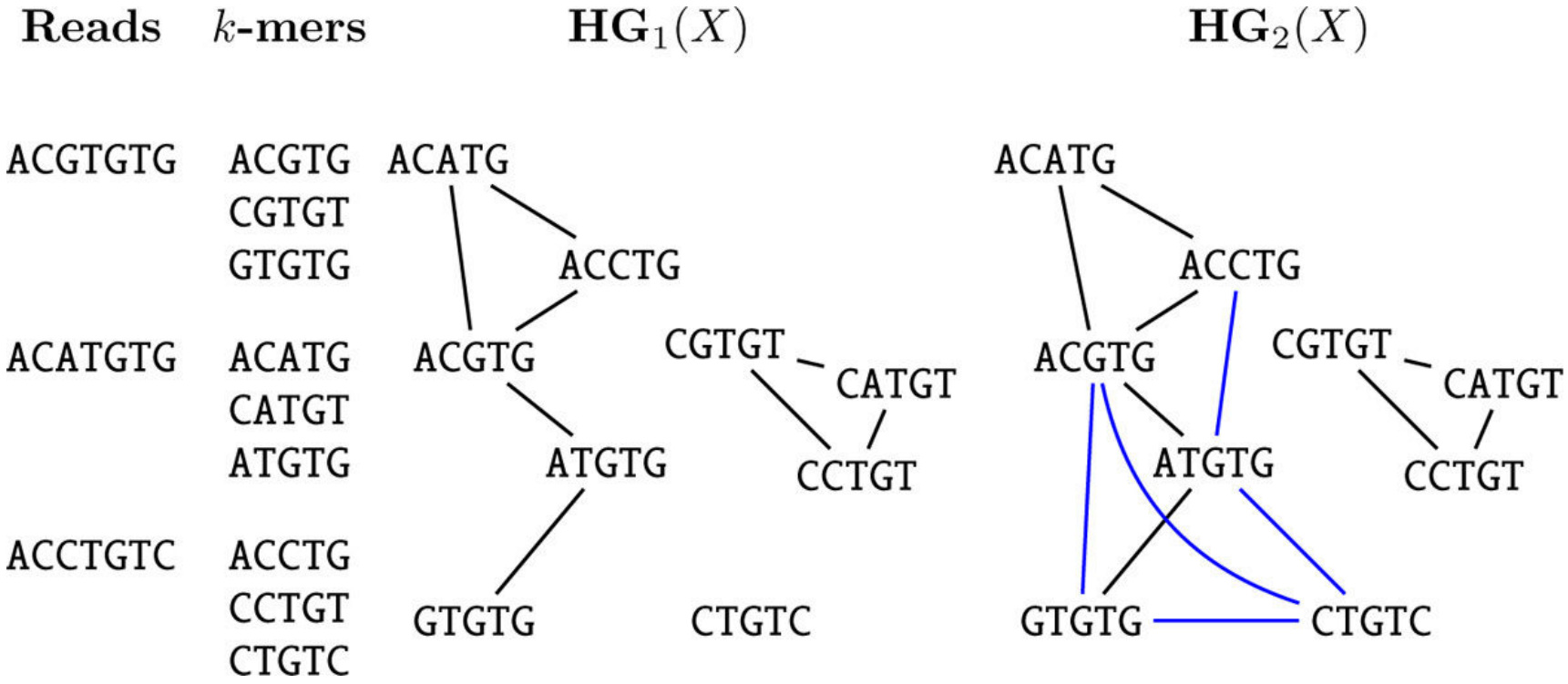


Hammer

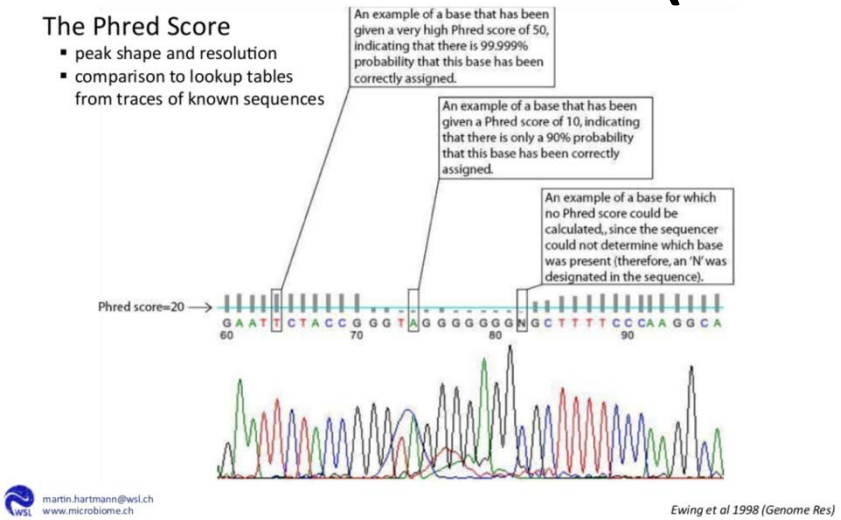
https://academic.oup.com/bioinformatics/article/27/13/i137/178096/Error-correction-of-high-throughput-sequencing

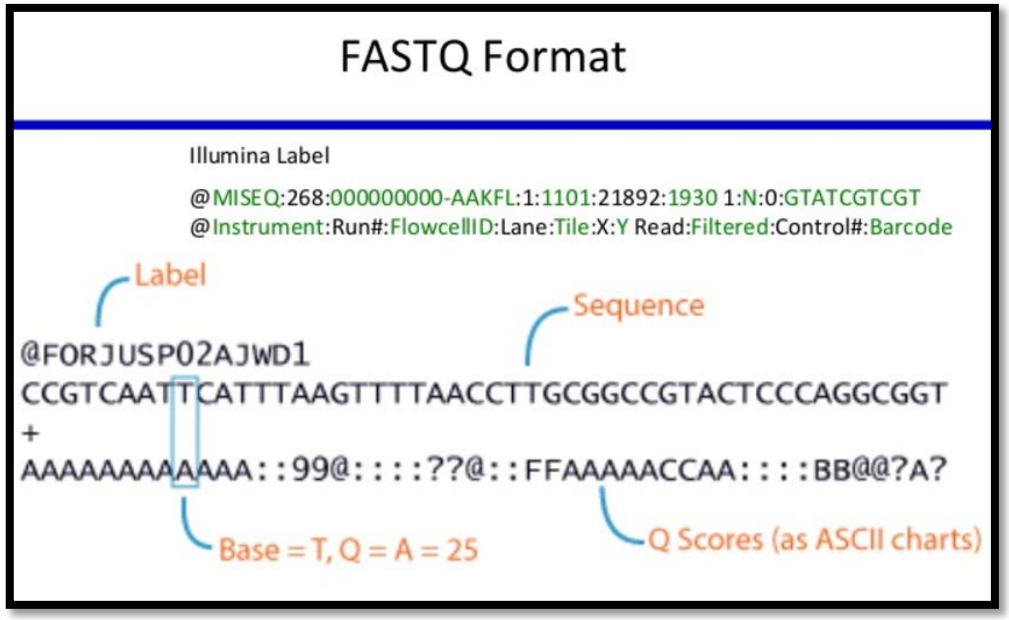
Correção de erros

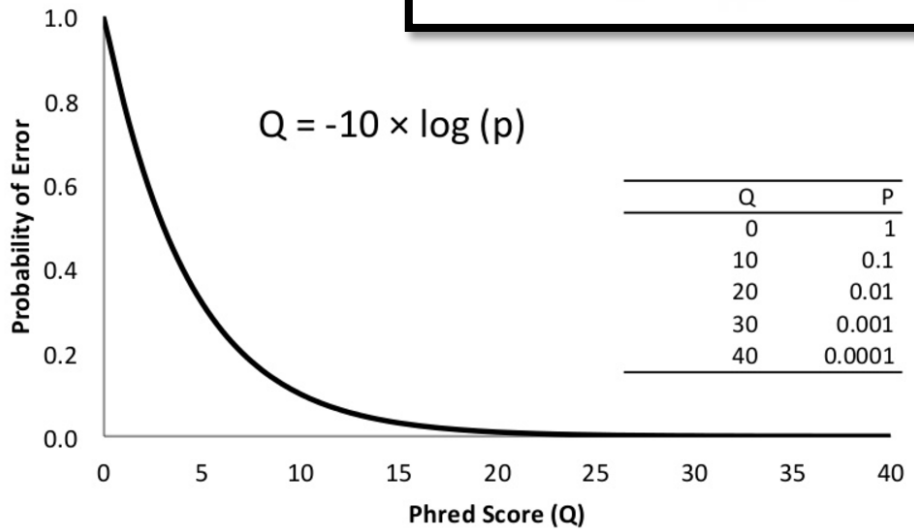
- Distância de Hamming
  - Número de posições em que as sequências divergem entre si.
- Grafo de Hamming
- Agrupamento



Qualidade de leitura das bases (Phred Score)







$$Q = -10 \times \log(p)$$

Q	P
0	1
10	0.1
20	0.01
30	0.001
40	0.0001

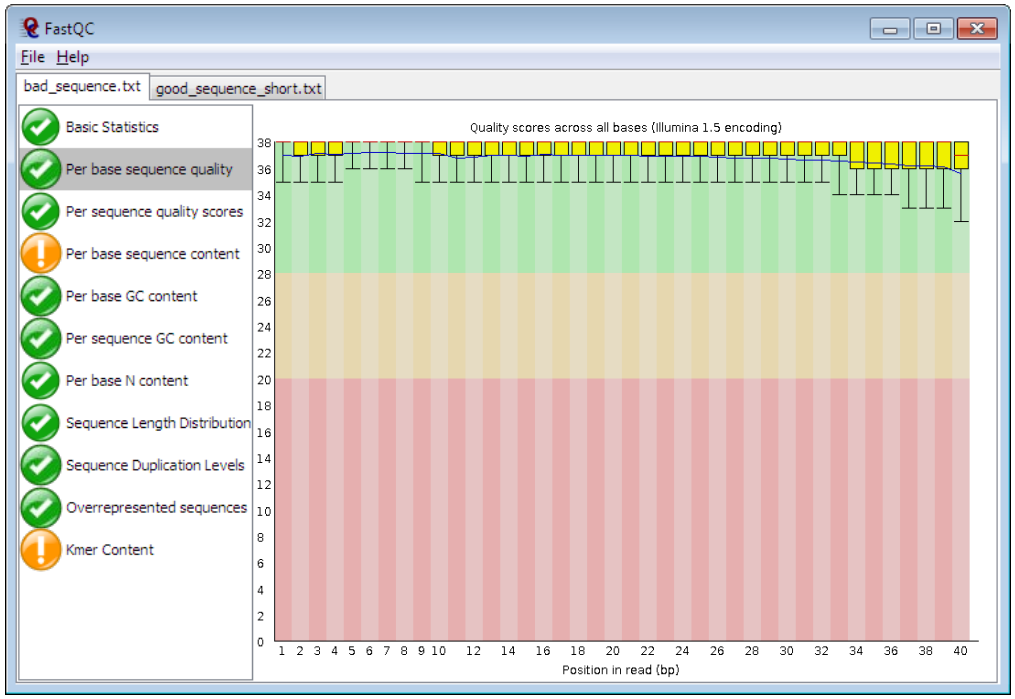
FastQC

Ferramenta para análise e controle de qualidade

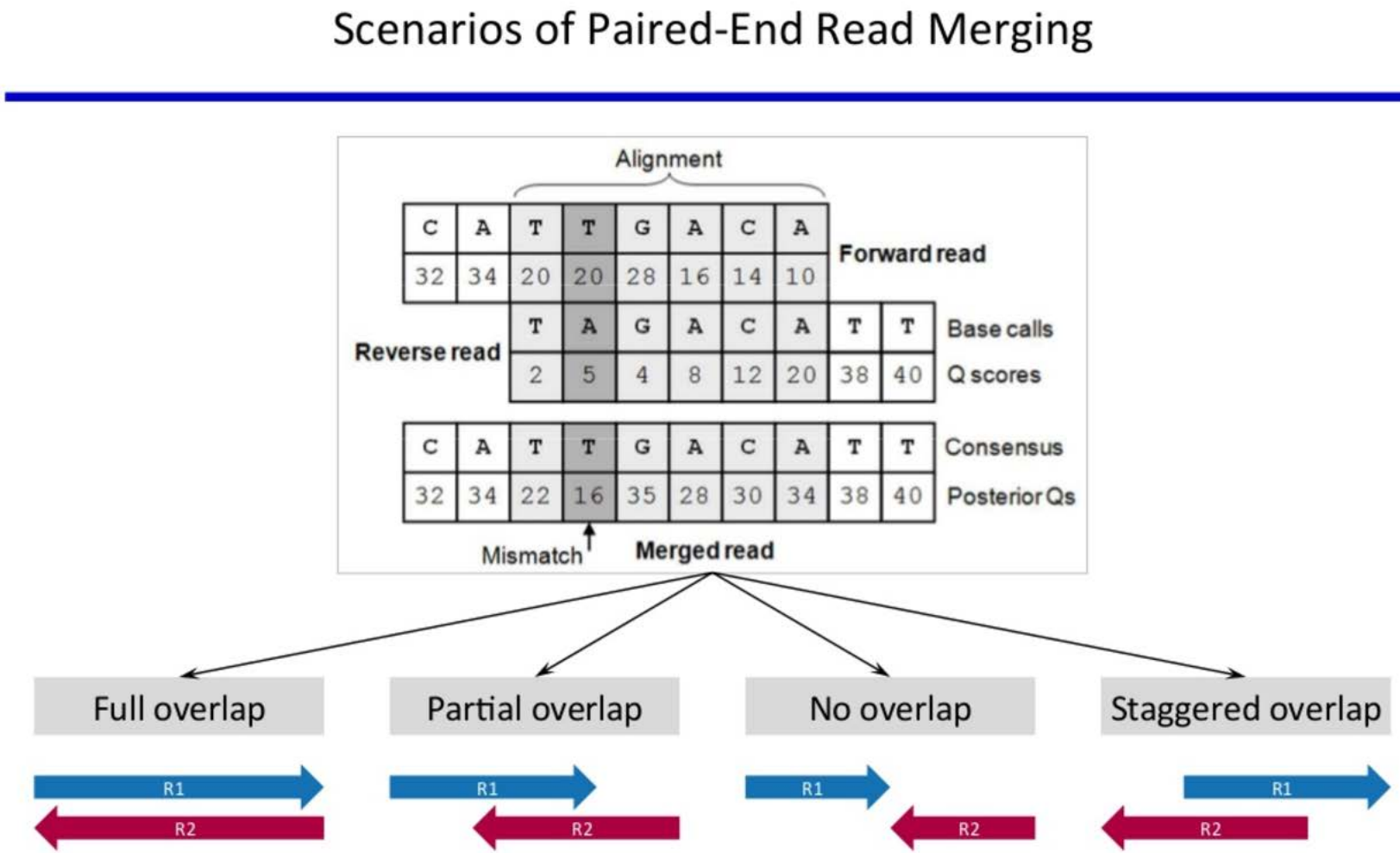
- http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/

fastqc seqfile1 seqfile2 .. seqfileN

fastqc [-o output dir] [--(no)extract] [-f fastq|bam|sam]
[-c contaminant file] seqfile1 .. seqfileN



Fusão de leituras *paired-end*



PEAR - Paired-End reAd mergeR

Bioinformatics. 2014 Mar 1;30(5):614-20. doi: 10.1093/bioinformatics/btt593. Epub 2013 Oct 18.

PEAR: a fast and accurate Illumina Paired-End reAd mergeR.

Zhang J[1], Kobert K, Flouri T, Stamatakis A.

Author information

Abstract

MOTIVATION: The Illumina paired-end sequencing technology can generate reads from both ends of target DNA fragments, which can subsequently be merged to increase the overall read length. There already exist tools for merging these paired-end reads when the target fragments are equally long. However, when fragment lengths vary and, in particular, when either the fragment size is shorter than a single-end read, or longer than twice the size of a single-end read, most state-of-the-art mergers fail to generate reliable results. Therefore, a robust tool is needed to merge paired-end reads that exhibit varying overlap lengths because of varying target fragment lengths.

RESULTS: We present the PEAR software for merging raw Illumina paired-end reads from target fragments of varying length. The program evaluates all possible paired-end read overlaps and does not require the target fragment size as input. It also implements a statistical test for minimizing false-positive results. Tests on simulated and empirical data show that PEAR consistently generates highly accurate merged paired-end reads. A highly optimized implementation allows for merging millions of paired-end reads within a few minutes on a standard desktop computer. On multi-core architectures, the parallel version of PEAR shows linear speedups compared with the sequential version of PEAR.

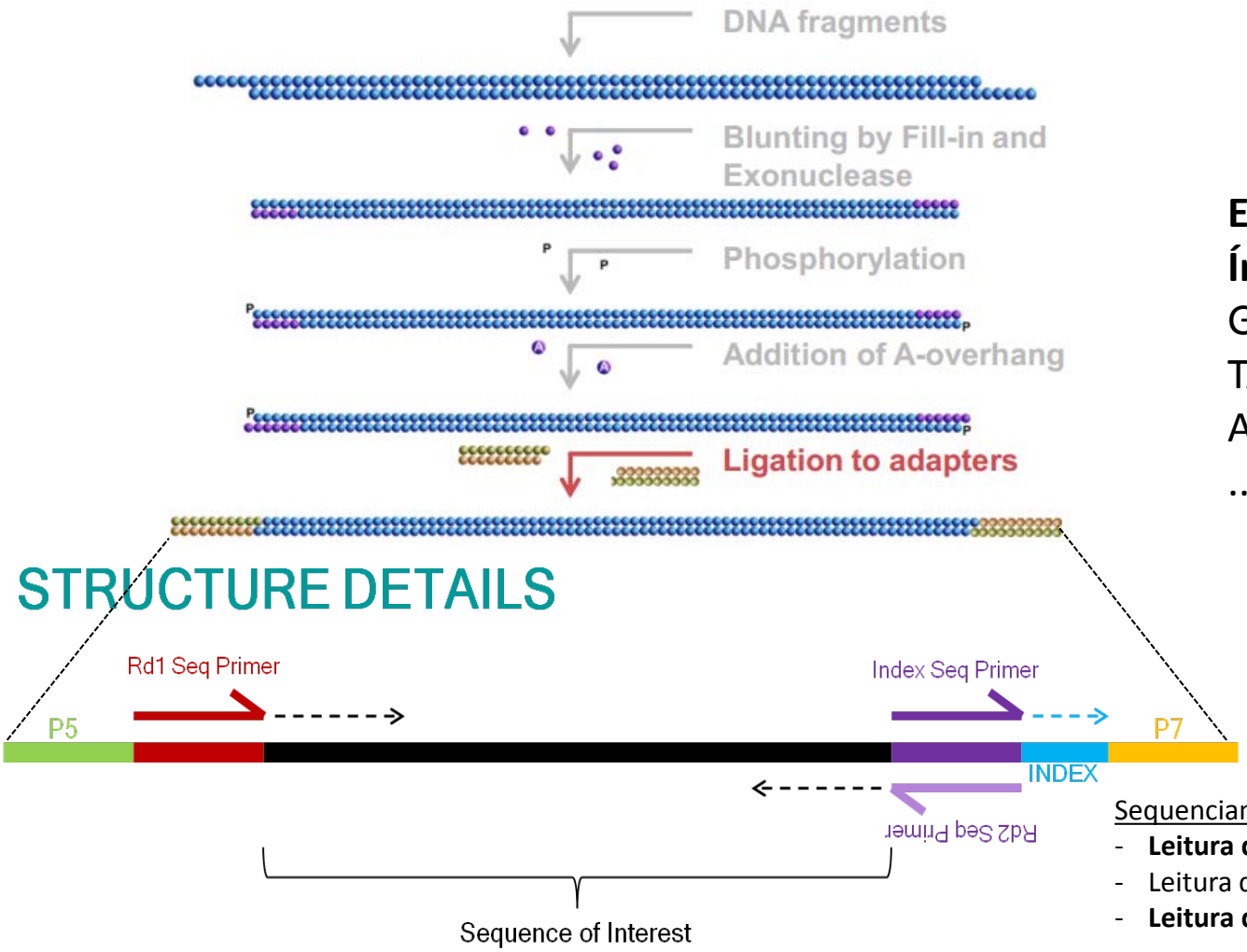
AVAILABILITY AND IMPLEMENTATION: PEAR is implemented in C and uses POSIX threads. It is freely available at http://www.exelixis-lab.org/web/software/pear.

PMID: 24142950 PMCID: PMC3933873 DOI: 10.1093/bioinformatics/btt593

[Indexed for MEDLINE] Free PMC Article

https://sco.h-its.org/exelixis/web/software/pear/

Estrutura dos fragmentos (poda de adaptadores)



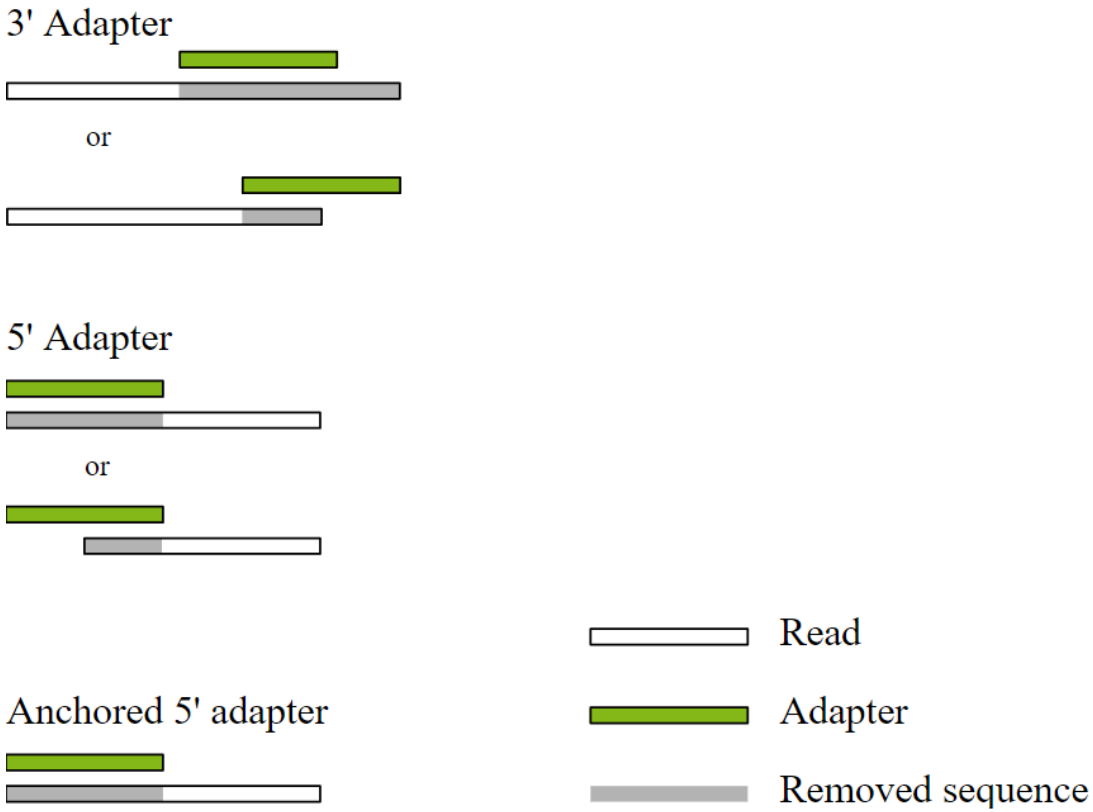
Exemplos de Índices:
GTGGCC
TAGCTT
ATTCCT
...

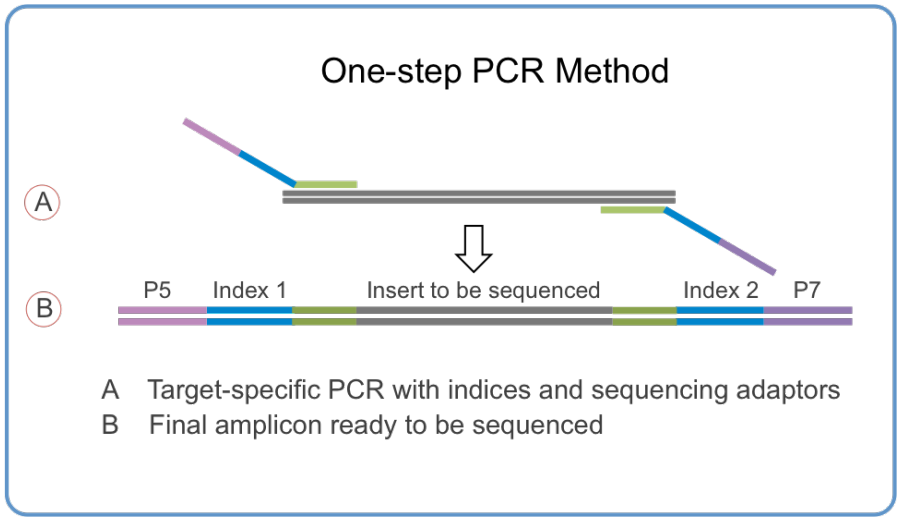
Sequenciamento em 3 etapas:
- **Leitura da extremidade P5;**
- Leitura do índice;
- **Leitura da extremidade P7;**

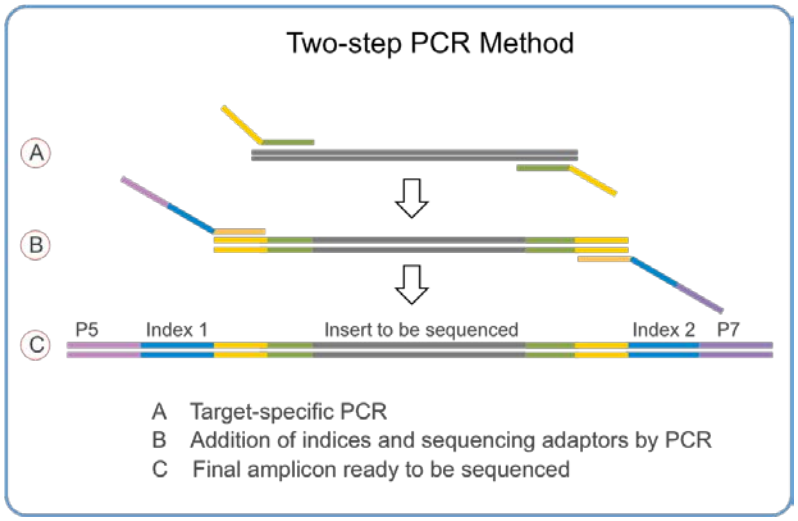
[http://nextgen.mgh.harvard.edu/CustomPrimer.html]

CutAdapt
(Poda – Adaptadores)

- https://github.com/marcelm/cutadapt







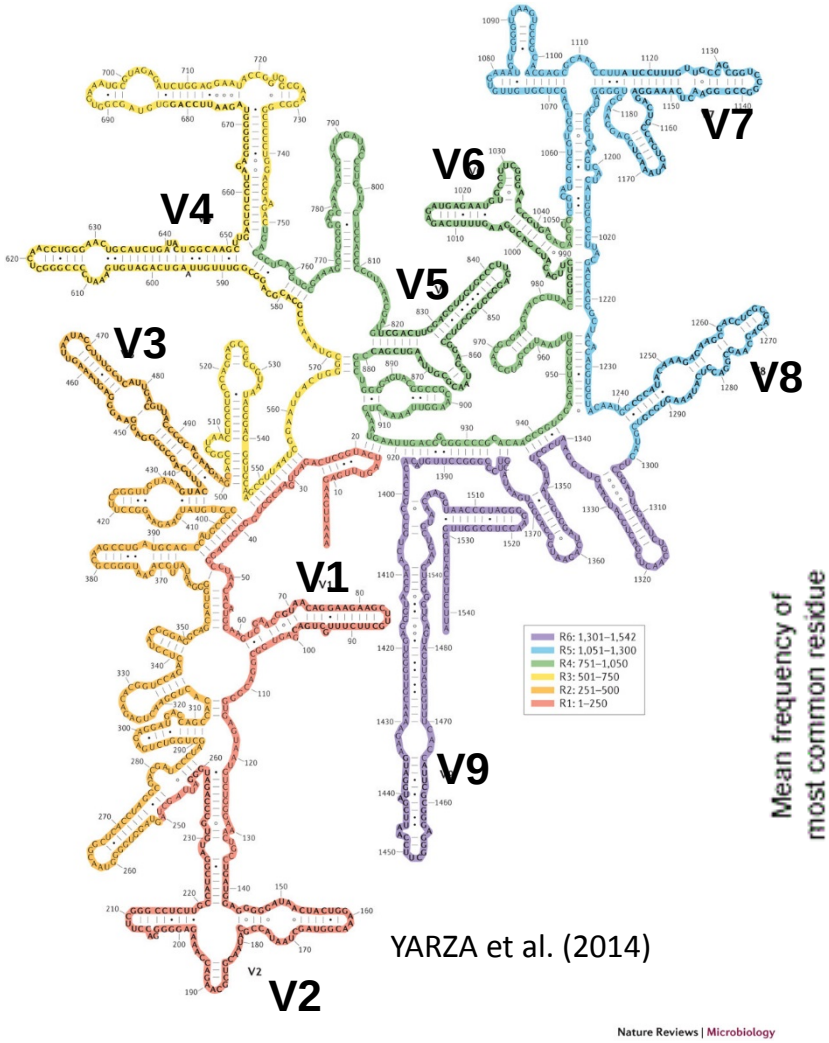
Introdução

ESTRATÉGIA BASEADA EM SEQUENCIAMENTO DE AMPLICONS ALVOS

Alvos

- Marcadores filogenéticos
  - Ex.: 16S (procariotos), ITS (eucariotos), …
- Marcadores funcionais
  - Ex.: nifH (fixação de nitrogênio atmosférico N_2 em por ex. amônia NH_3)

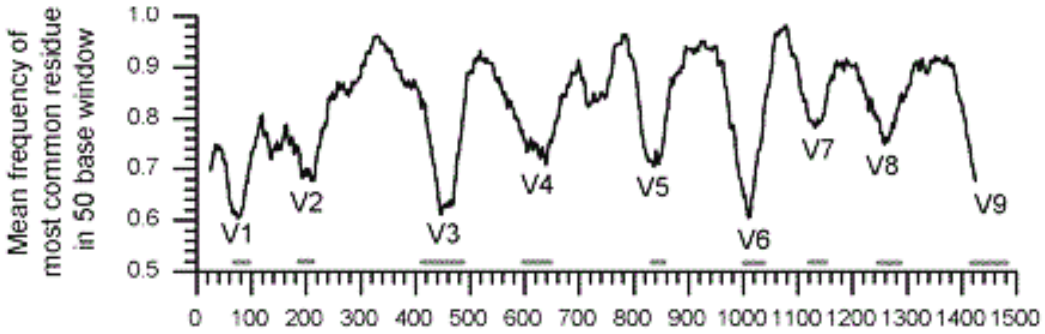
O gene 16S rRNA



YARZA et al. (2014)

baseado em *E. coli*

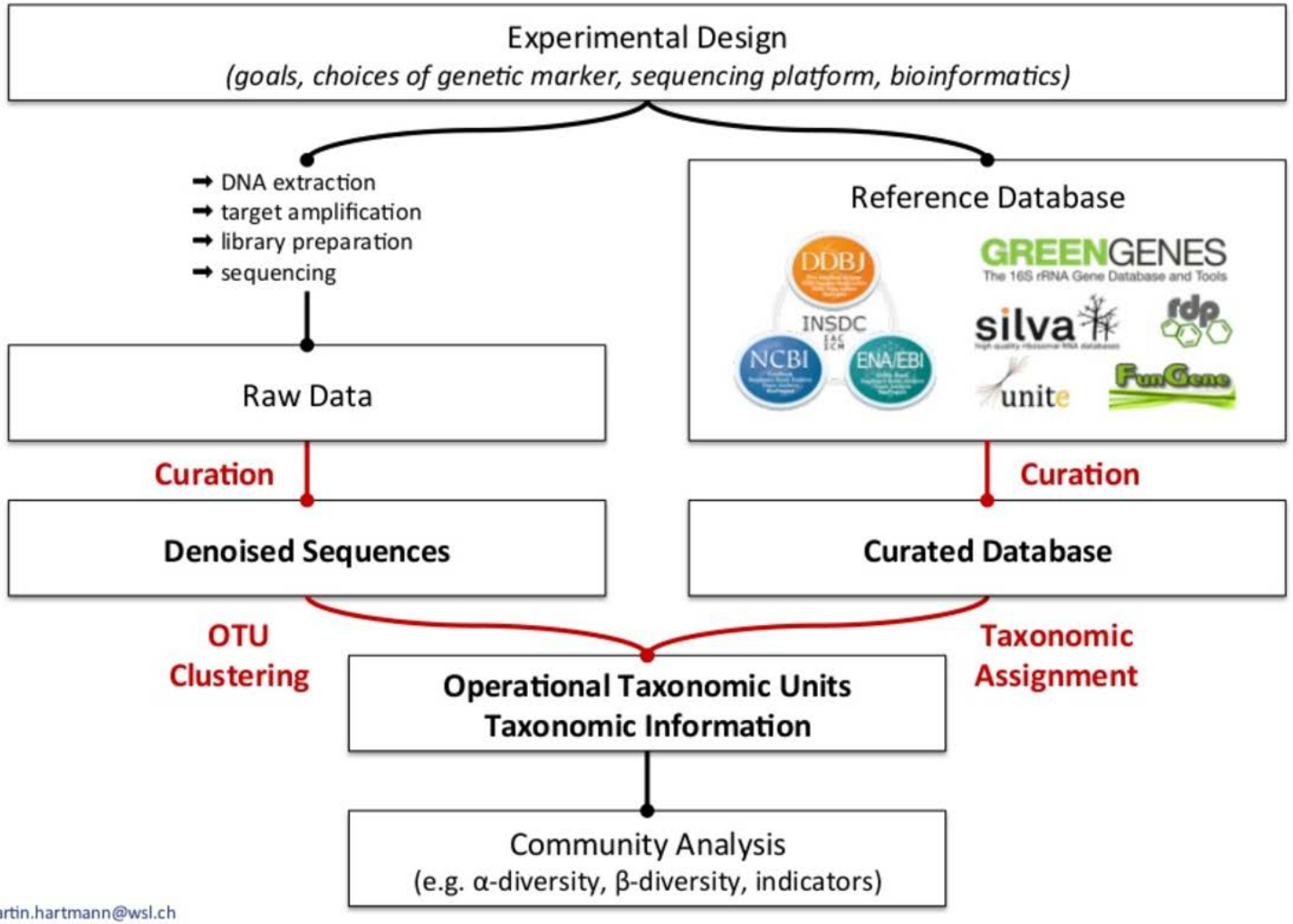
Regiões Hipervariáveis



Ashelford et al. (2005)

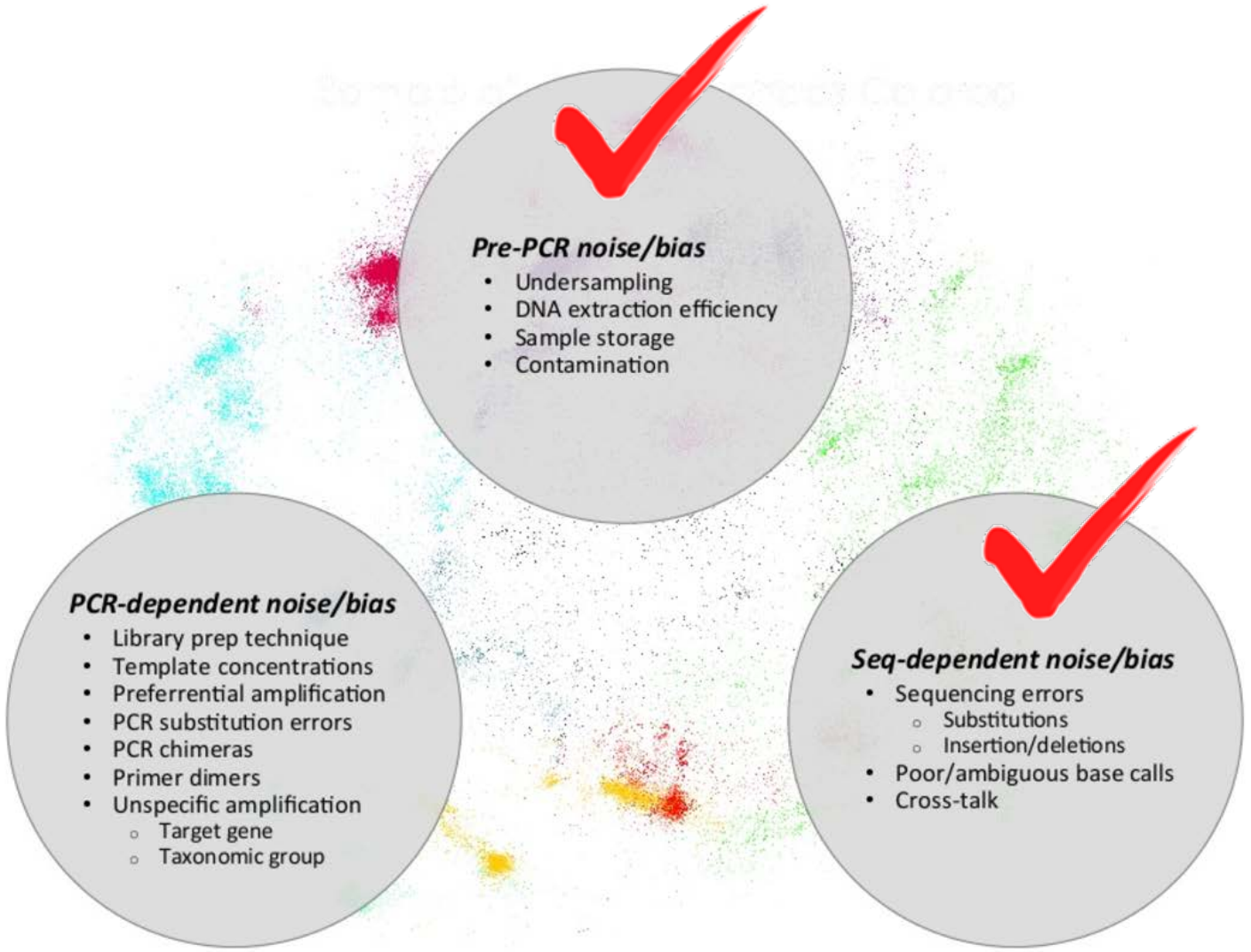
"O gene 16S rRNA baseado em E. coli" (YARZA et al., 2014). DOI: 10.1038/nrmicro3330. (WANG et al., 2007). DOI: 10.1128/AEM.00062-07. (Ashelford et al. 2005). DOI: 10.1128/AEM.71.12.7724-7736.2005

Workflow básico



martin.hartmann@wsl.ch
www.microbiome.ch

Fontes de ruídos



cross-talk ocorre quando uma leitura é atribuída a uma amostra incorretamente (sequenciamento multiplex)

martin.hartmann@wsl.ch
www.microbiome.ch

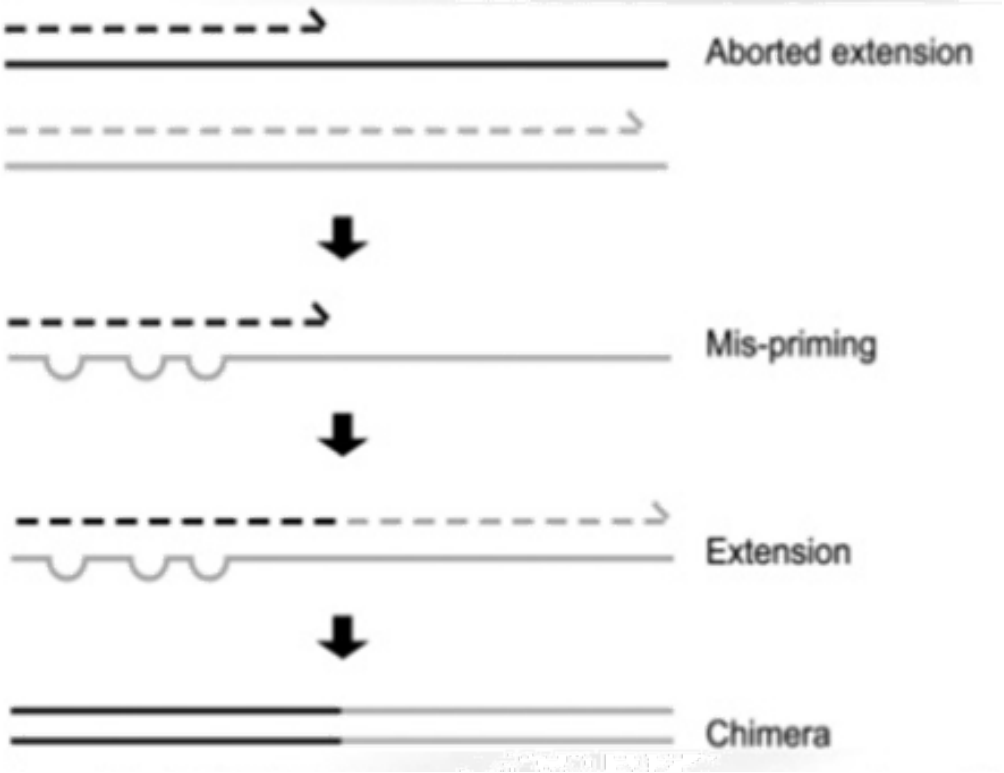
credit: http://salsahpc.indiana.edu/millionseq/

Formação de quimeras

1) Extensão de *primer* abortada e formação de novo *primer*

2) Anelamento desse novo *primer* em outra sequência de espécie diferente

3) Extensão desse novo primer e formação de sequências quimeras (as quais serão amplificadas) nos ciclos seguintes



Hass B.J. et al (2011) *Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons*, Genome Res. 21: 494-504.

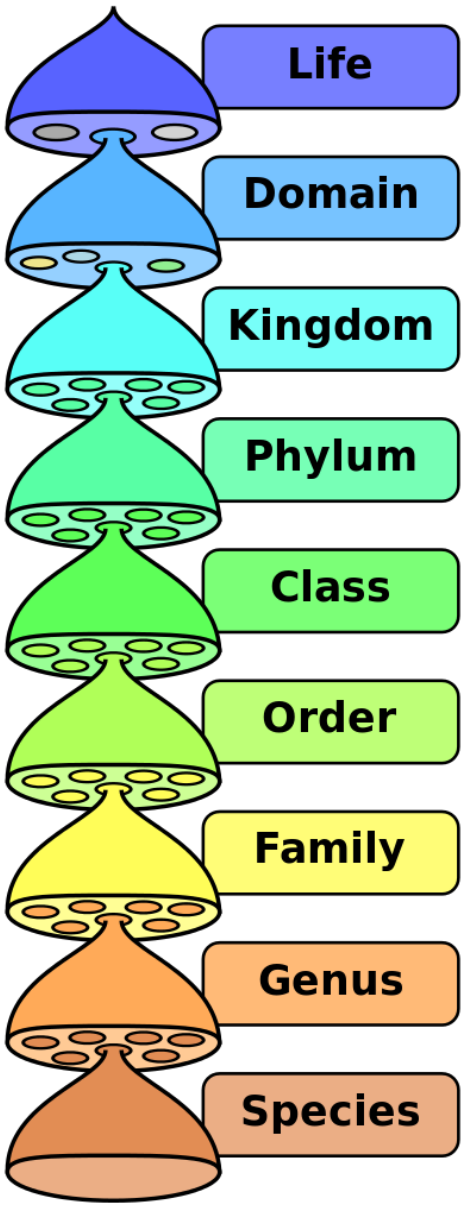
Identificação de quimeras (UCHIME)

Parsimony model

```
A       CCTTGGTAGGCCGtTGCCCTGCCAACTAGCTAATCAGACGCgggtCCATCtcaCACCaccggAgtTTTtcTCaCTgTacc
Q       CCTTGGTAGGCCGCTGCCCTGCCAACTAGCTAATCAGACGCATCCCCATCCATCACCGATAAATCTTTAATCTCTTTCAG
B       TCTTGGTgGGCCGtTaCCCcGCCAACaAGCTAATCAGACGCATCCCCATCCATCACCGATAAATCTTTAAaCTCTTTCAG
Diffs   A      A     p A   A      A                BBBB     BBB    BBBBB BB   BBa B  B BBB
Votes   +      +     0 +   +      +                ++++     +++    +++++ ++   ++! +  + +++
Model   AAAAAAAAAAAAAAAAAAAAAAAAAxxxxxxxxxxxxxxBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
```

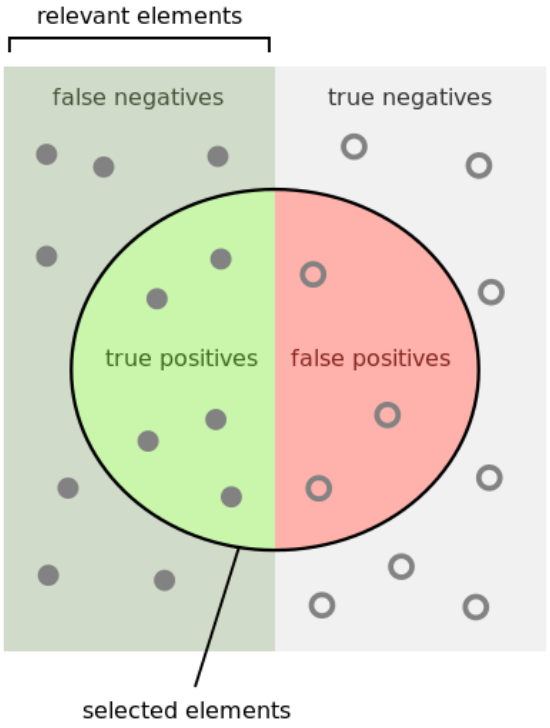
Binning

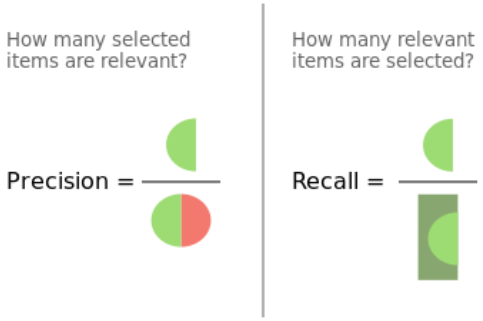
- Em metagenômica, é o processo de **agrupar *reads* ou *contigs* e atribuir a ele uma Unidade Taxonômica Operacional** (*Operational Taxonomic Unit* – OTU)
  - Agrupamento em OTUs (*clustering*) – critério usual 97% similaridade
    - Melhor aproximação de espécie
      - Não é possível lidar com reads livres de erros
      - Não é possível identificação sempre ao nível de espécie
      - Variabilidade intra-espécie



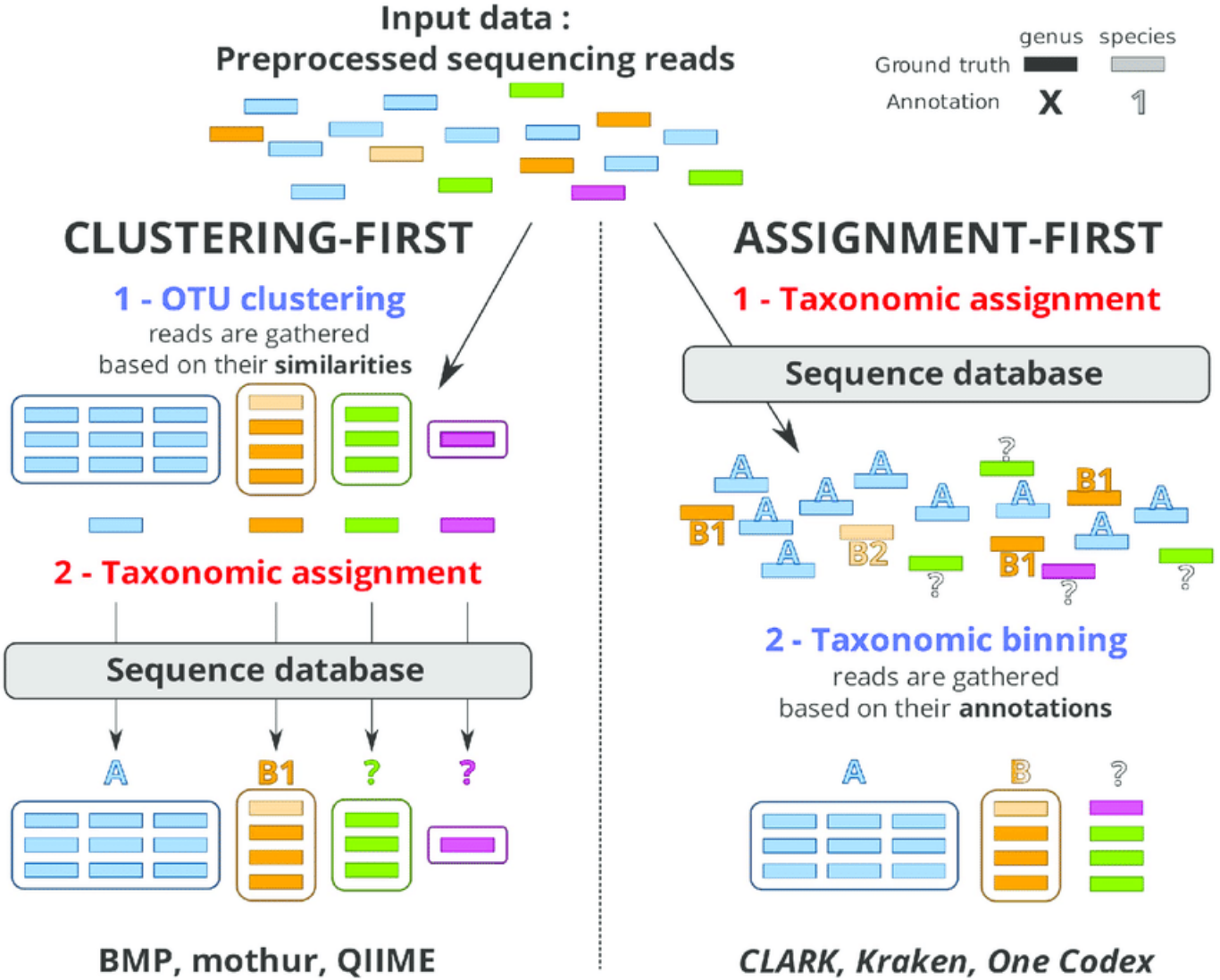
Remoção de *singleton reads*

- *Singleton reads* – leituras que aparecem uma única vez
  - Alguns *singletons* possem mais de 3% de divergência e formam OTUs espúrias
  - Sugestão
    - Remover os *singletons* antes do agrupamento (*clustering*) e mapeá-los posteriormente
  - Remoção
    - Aumenta a especificidade ao custo de uma pequena perda de sensibilidade



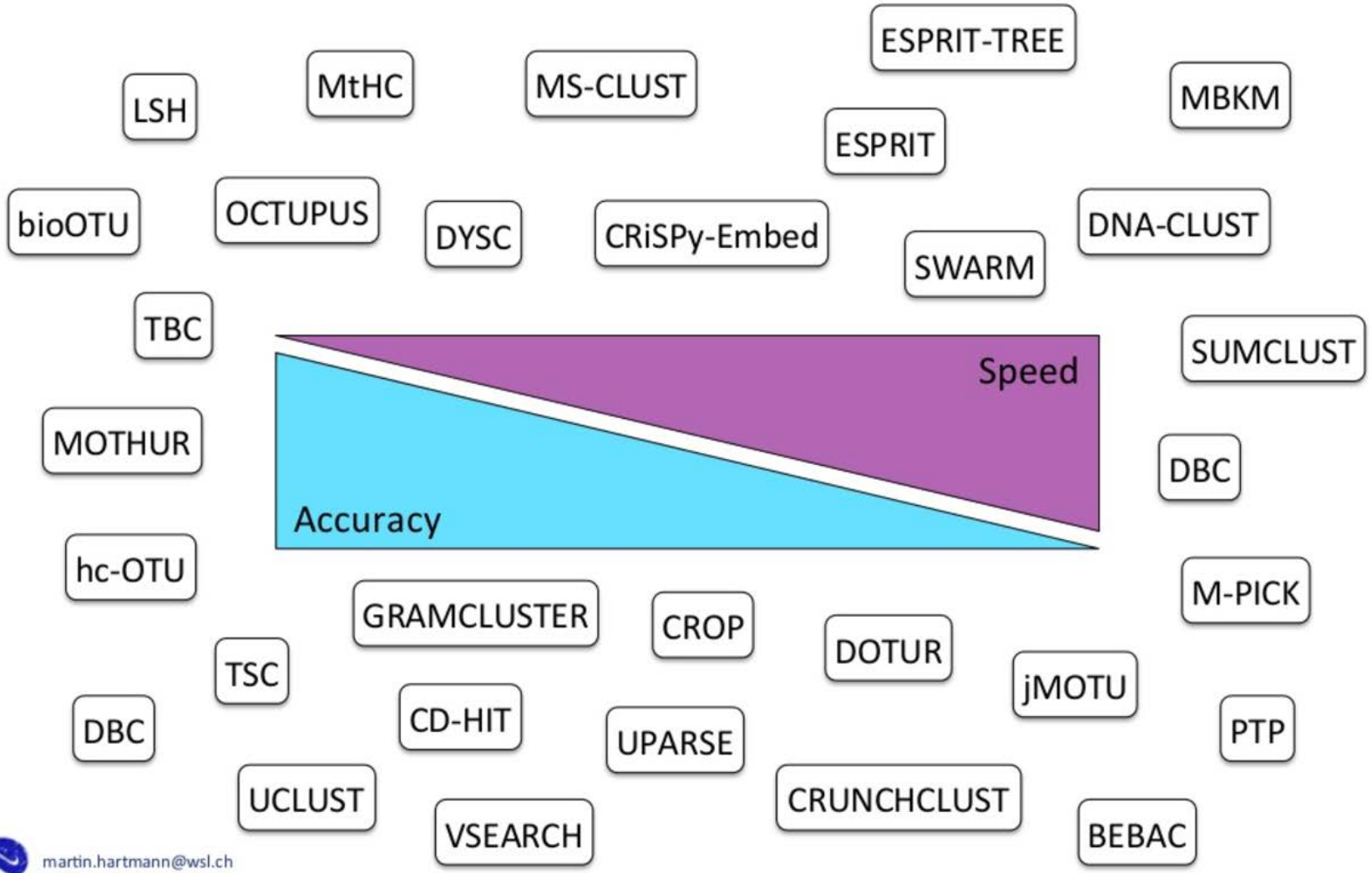


Estratégias de Identificação Taxonômica



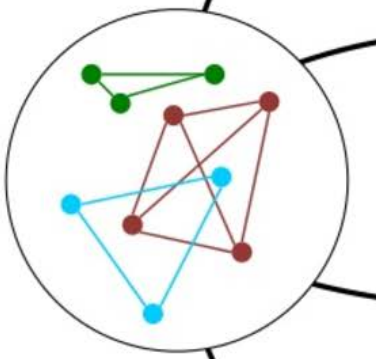
DOI: 10.1371/journal.pone.0169563

Programas para Agrupamento (*clusterina*)



martin.hartmann@wsl.ch
www.microbiome.ch

Clustering



OTU Definition
- Based on taxonomy
- Based on sequence similarity
- Based on phylogenetic tree

Clustering Strategy
- De novo clustering
- Closed-reference clustering
- Open-reference clustering

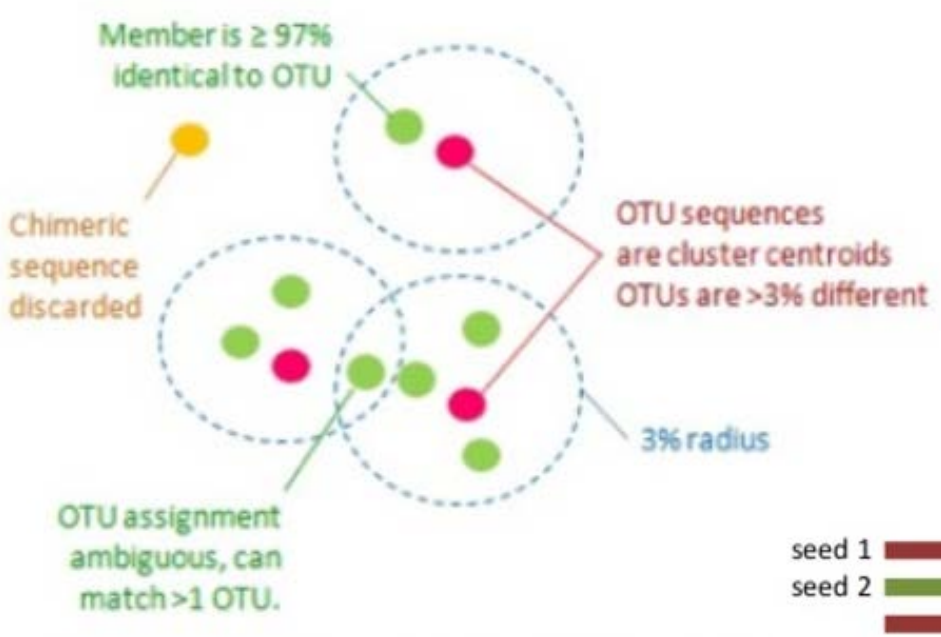
Distance measurement
- Multiple sequence alignment
- Pairwise sequence alignment
- Alignment-free

Algorithm
- Hierarchical clustering
- Seed-Based clustering
- Model-based clustering

martin.hartmann@wsl.ch
www.microbiome.ch

UPARSE

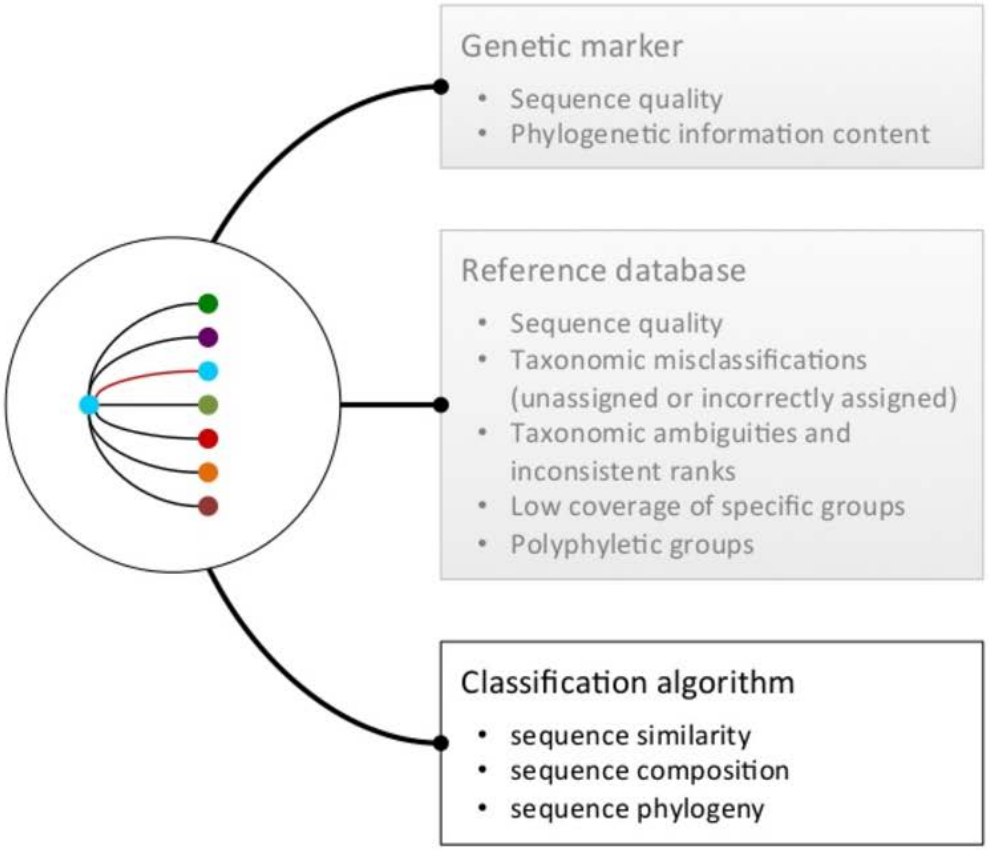
Seed-based (greedy) clustering



As *reads* são ordenadas (por abundância) e comparadas entre si, sendo possíveis dois casos:

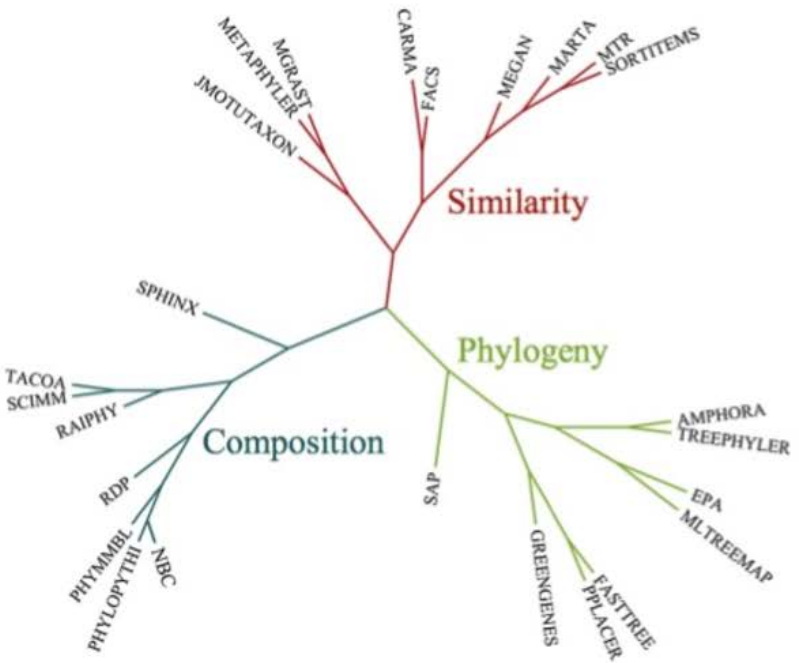
1) similaridade >= 97% - membro do *cluster* com centroid mais similar e mais abundante
2) similaridade < 97% - nova *seed*

Identificação Taxonômica



martin.hartmann@wsl.ch
www.microbiome.ch

Algoritmos para Identificação



Similarity-based

Find homology or minimum alignment distance

Tools:
- local alignments (e.g. BLAST, MEGAN, METAXA2, RTAX)
- global alignments (e.g. GAST)
- overlap alignments (e.g. SINA)

Pro/Con:
- good accuracy for similar sequences
- performs less well on distant lineages
- can be slow on large reference databases

Composition-based

Detect specific features

Tools:
- kmer searches (e.g. NBC/RDP, UTAX, SINTAX)
- hidden Markov models (e.g. PHYMMBL, C16S)

Pro/Con:
- computationally efficient and fast
- performs well on distant lineages
- training required
- limited resolution for shorter sequences

Phylogeny-based

Evolutionary model to determine best placement

Tool:
- ML, NJ, Bayesian (e.g. PPLACER, EPA)

Pro/Con:
- great accuracy for similar sequences
- classification in its evolutionary context
- computationally complex
- requires accurate reference tree
- difficult for non-coding regions

Bazinet & Cummings 2012 (BMC Bioinformatics)

Conclusions

- To date, no algorithm is convincingly outperforming the others.
- Other factors such as genetic marker/locus, sequence quality, and integrity of reference database are likely more crucial.

martin.hartmann@wsl.ch
www.microbiome.ch

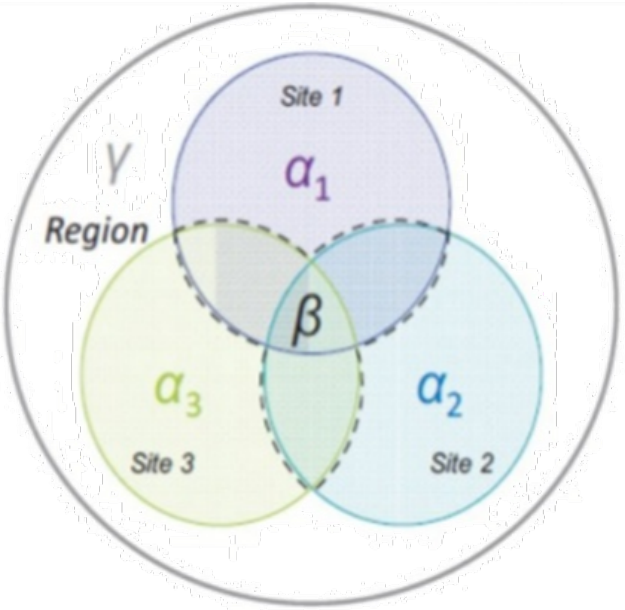
Análise das Comunidades

- Diversidade

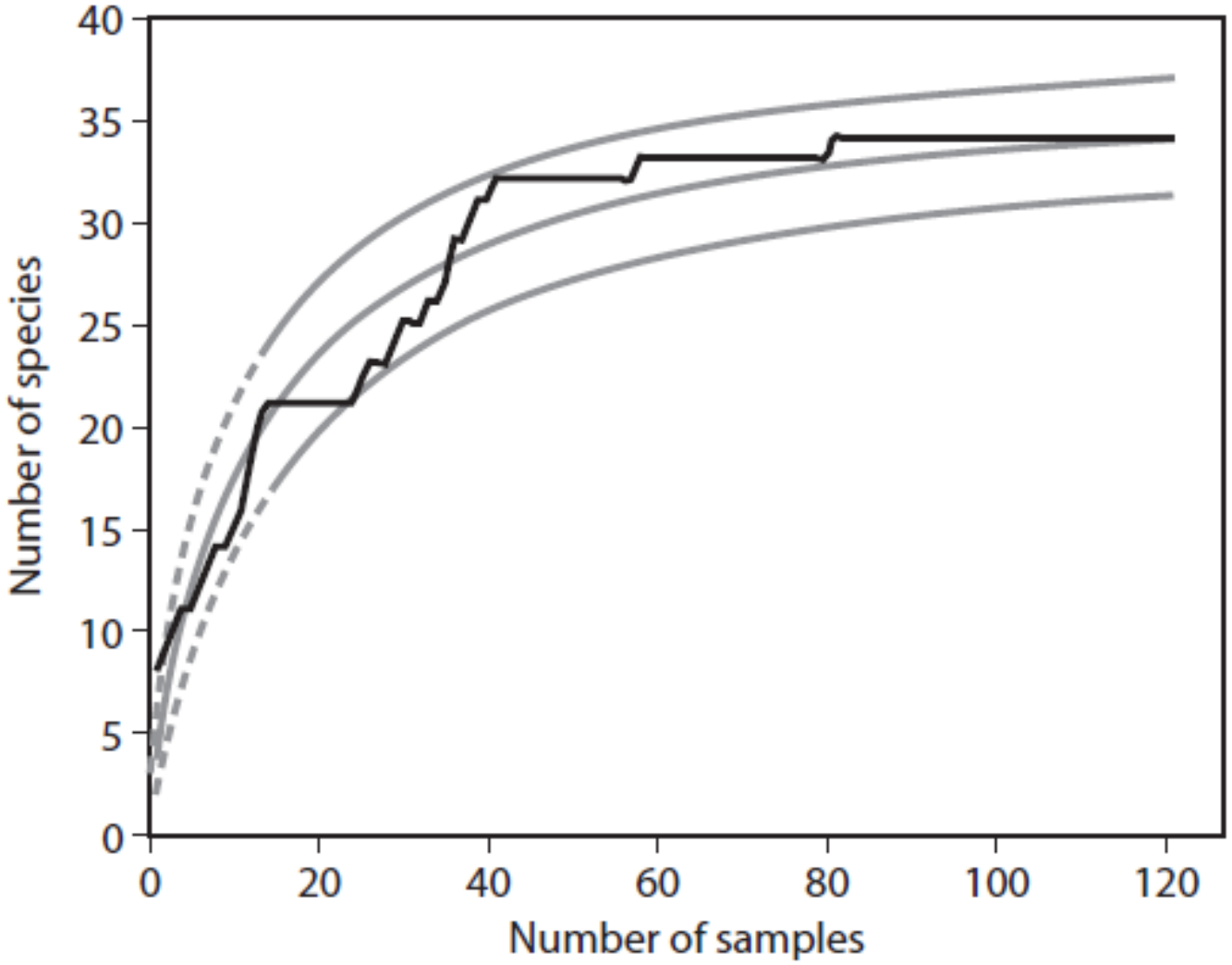
α-diversidade (alfa): diversidade de uma amostra/bioma (sensível a delimitação de ambiente e como se realiza a amostragem);

β-diversidade (beta): **diversidade entre habitats**, influenciado pela heterogeneidade da estrutura das comunidades (composição e proporção da espécies).

γ-diversidade (gama): **diversidade regional**, relacionada ao número total de espécies observado em todos os habitats dentro de uma área geográfica.



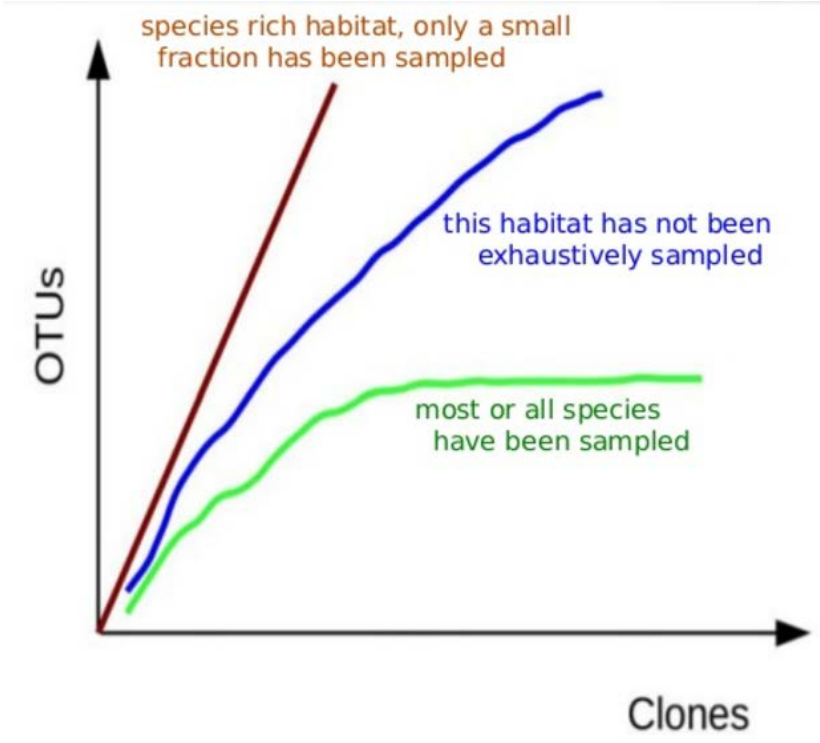
Curva de rarefação

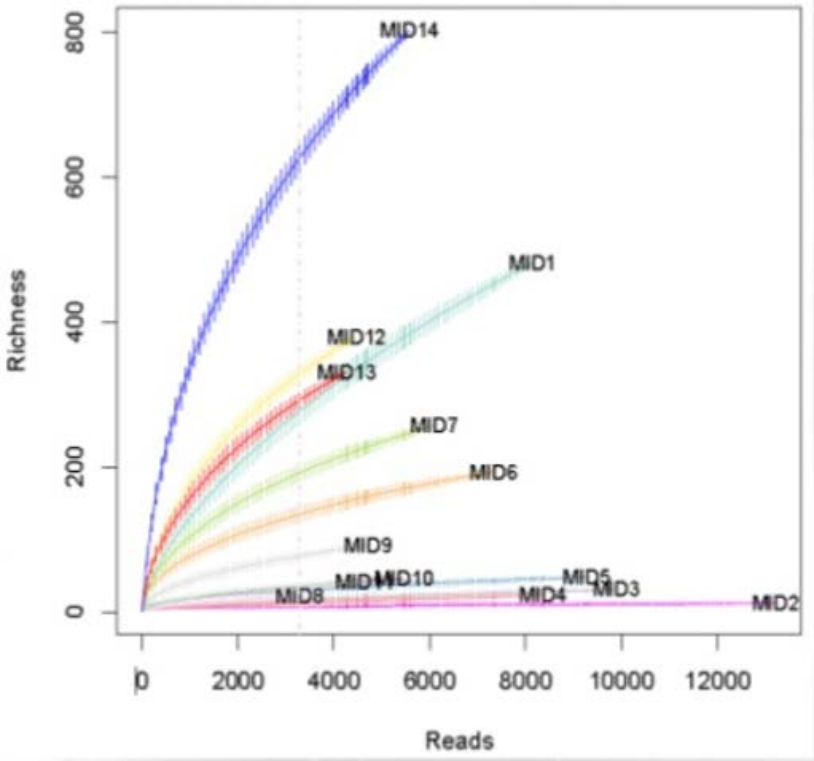


Curva de rarefação

Curva de acumulação de espécies ou curva coletora

Avaliação da amostragem





Medidas e estimativas da riqueza de espécies

$$S_{est} = S_{obs} + \frac{a^2}{2b}.$$

S_{est} = Riqueza estimada

S_{obs} = Espécies observadas

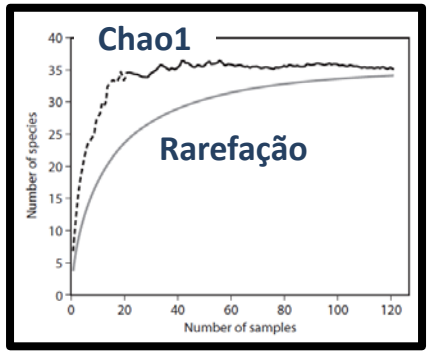
a = espécies contendo único indivíduo – *singletons*

b = espécies contendo dois indivíduos – *doubletons*

S_{obs} = 34 espécies
a = 2
b = 2

$$S_{est} = 34 + \left(\frac{2^2}{2 \cdot 2}\right)$$

=> S_{est} = 35 espécies



A riqueza aumenta com o acréscimo de espécies raras.

" Medidas e estimativas da diversidade de espécies" Colwell RK (2009) Biodiversity: concepts, patterns and measurement. In SA Levin. The Princeton guide to ecology. Princeton, NJ, USA: Princeton University Press. pp. 257–263.

Equitatividade

- Em Ecologia, é o termo empregado para definir a **uniformidade**, ou homogeneidade, da distribuição de abundância de espécies em uma comunidade.

- Em uma comunidade, a equitatividade será **baixa** quando há **poucas espécies altamente dominantes** em meio a um grande número de espécies raras. Se não houver espécies altamente dominantes, a equitatividade será maior.

- Geralmente é expressa de forma numérica (variando de zero a 1), derivada de algum índice de diversidade específico.

Medidas e estimativas da diversidade de espécies

$$H = -\sum_{i=1}^{s} p_i \ln p_i$$

Diversidade Shannon

$$D = 1 - \sum_{i=1}^{s} p_i^2$$

Diversidade Simpson

Combinam riqueza e equitatividade em uma única medida

" Medidas e estimativas da diversidade de espécies" Colwell RK (2009) Biodiversity: concepts, patterns and measurement. In SA Levin. The Princeton guide to ecology. Princeton, NJ, USA: Princeton University Press. pp. 257–263.

Medidas e estimativas da diversidade de espécies

$$H = -\sum_{i=1}^{s} p_i \ln p_i$$

Diversidade Shannon

$$D = 1 - \sum_{i=1}^{s} p_i^2$$

Diversidade Simpson

No. Indivíduos	pi	lnpi	pi*lnpi
10	0,10	-2,30	-0,23
10	0,10	-2,30	-0,23
10	0,10	-2,30	-0,23
10	0,10	-2,30	-0,23
10	0,10	-2,30	-0,23
10	0,10	-2,30	-0,23
10	0,10	-2,30	-0,23
10	0,10	-2,30	-0,23
10	0,10	-2,30	-0,23
10	0,10	-2,30	-0,23
N=100	**S=10**	**H'=**	**2,30**

No. Indivíduos	pi		pi^2
10	0,1		0,01
10	0,1		0,01
10	0,1		0,01
10	0,1		0,01
10	0,1		0,01
10	0,1		0,01
10	0,1		0,01
10	0,1		0,01
10	0,1		0,01
10	0,1		0,01
N=100	**S=10**	**D=**	**0,9**

" Medidas e estimativas da diversidade de espécies" Colwell RK (2009) Biodiversity: concepts, patterns and measurement. In SA Levin. The Princeton guide to ecology. Princeton, NJ, USA: Princeton University Press. pp. 257–263.

qiime

- http://qiime.org/
- *QIIME is an **open-source bioinformatics pipeline** for performing **microbiome analysis** from **raw DNA sequencing data**. QIIME is designed to take users from raw sequencing data generated on the **Illumina or other platforms** through publication quality graphics and statistics. This includes **demultiplexing and quality filtering, OTU picking, taxonomic assignment, and phylogenetic reconstruction, and diversity analyses and visualizations**. QIIME has been applied to studies based on billions of sequences from tens of thousands of samples.*

dgpinheiro@gmail.com